Information extraction in the field of chemistry: a review

1. Introduction（陈辉）

The contemporary landscape of modern chemistry and materials science research is marked by an exponential growth in the volume of scientific literature. This vast repository of knowledge, which encompasses detailed information on novel compounds, properties, synthesis routes, and experimental results, is predominantly stored in unstructured natural language text within journal articles, patents, and reports. While this format is well-suited for human communication, it poses a significant challenge for data-driven research approaches that rely on structured, machine-readable data. Consequently, transforming this extensive pool of unstructured text into structured, actionable knowledge is crucial for accelerating discovery and innovation in the domains of chemistry and materials science.

Historically, the challenge of extracting information from chemical literature has been addressed through the application of Natural Language Processing (NLP) techniques. Early efforts in this field leveraged methods adapted from general NLP, such as rule-based systems and statistical models, to identify entities (e.g., compounds, properties, and synthesis procedures) and their interrelationships. These approaches, including Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE) pipelines, have demonstrated satisfactory performance for specific, narrowly defined problems and datasets. Various traditional NLP techniques, such as dictionary look-ups, rule-based methods, semi-supervised approaches, and machine learning models utilizing architectures like BiLSTM-CRF for NER and relation extraction, have been developed for chemical information extraction. Pre-trained word embeddings, such as Word2vec and GloVe, and later contextual embeddings derived from models like BERT, have played a pivotal role in capturing the semantic meaning of chemical language. Although previous reviews have documented the progress of NLP and text mining in this domain, the rapid evolution of NLP techniques, particularly the advent of Large Language Models (LLMs), necessitates a comprehensive overview of the latest and most advanced methods employed for information extraction in chemistry and materials science. This review aims to provide a timely and structured analysis of these developments.

In recent years, the field of NLP has been profoundly transformed by the emergence of Large Language Models (LLMs) such as GPT, Falcon, and BERT, which are built upon the Transformer architecture. Trained on vast amounts of text data, LLMs exhibit remarkable general "intelligence" capabilities, including advanced text understanding and generation. The success of LLMs presents a significant opportunity and holds immense potential for information extraction in the chemical and materials science domains. These models overcome some limitations of traditional methods by enabling context-aware interpretation of unstructured literature and flexible entity recognition. LLMs can redefine extraction tasks, often framing them as text generation problems, and techniques like prompt engineering offer a novel way to guide these models for precise and relevant information extraction. Moreover, domain-specific LLMs fine-tuned on chemical or materials science corpora, such as SciBERT, MatSciBERT, or SteelBERT, have demonstrated enhanced performance in domain-specific NLP tasks, including information extraction. These advancements enable researchers to benefit from more accurate information extraction and improved understanding of complex chemical concepts.

This review provides a comprehensive overview of recent advancements in applying NLP, with a particular focus on Large Language Models, for information extraction in chemistry and materials science. We will summarize the evolution of NLP methods in this field, encompassing both traditional pipelines and modern LLM-based approaches. We will detail how these techniques are applied to extract various types of chemical information, such as composition, properties, and synthesis procedures. Furthermore, we will consolidate and discuss the latest and most advanced methodologies and techniques currently employed for information extraction, drawing insights from recent research. Finally, we will explore the future potential of LLMs in this domain and outline key challenges and opportunities that will shape the landscape of chemical information extraction in the coming years.

介绍太长了，对于规则系统、传统方法的介绍待删减，着重强调LLM in chemistry的潜力

2. NER+chemistry（陈辉）

a. Domain-level tasks

In the fields of chemistry and materials science, Named Entity Recognition (NER) is widely applied to extract specific entities from unstructured text, supporting the creation of structured datasets for further analysis. These domain-level tasks are highly specialized to identify key components, processes, and properties relevant to each area. For instance, in the synthesis of inorganic materials, NER is used to extract detailed procedures and conditions. In the study of advanced materials like metallic glasses and Metal-Organic Frameworks (MOFs), NER targets properties and structural information. Additionally, in the battery domain, NER is employed to categorize components such as anode and cathode materials. The following table provides a comprehensive overview of these tasks and the entities targeted for extraction.

这里简要的举例还是替换成xx学者的工作比较好，之后进行替换

表格是否全面，还需要进一步梳理

Domain-level Task	Targeted Entities

General Materials Information Extraction	Compounds, Alloy compositions, Properties, Synthesis actions, Parameters
Synthesis Procedure Extraction (General Inorganic Materials)	Targets, Precursors, Operations, Conditions, Reactions
Synthesis and Processing Extraction (Alloys, specifically Superalloys)	Synthesis and processing actions, Chemical compositions, Parameters
Materials Properties Extraction (Specific Materials like Metallic Glasses, Ternary Chalcogenides)	Material, Value, Unit (e.g., bulk modulus, critical cooling rates)
Metal-Organic Frameworks (MOFs) and Reticular Materials Information Extraction	Synthesis details, Chemical formulae, Applications, Further descriptions
Chemical Reaction Extraction	Reactants, Products, Reagents, Solvents, Catalysts, Conditions
Battery Device Component Categorization	Anode materials, Cathode materials, Electrolyte materials
Organic Field-Effect Transistors (OFETs) Experimental Parameter Extraction	Experimental parameters for OFETs
Polymer Synthesis Routes	Polymer synthesis routes
Catalysis Information Extraction	Catalyst components, Catalyst characteristics
NMR Data Extraction	Specific entities not detailed in sources

b. Data annotation

In the realm of chemical and materials science, preparing data for Named Entity Recognition (NER) tasks often begins with sourcing relevant literature. For instance, researchers like Kononova et al.9 have utilized repositories such as EuroPMC55 and arXiv57, which offer a wealth of open-access scientific articles. These platforms provide APIs for automated access, facilitating the collection of large datasets.

Processing data for NER tasks in chemistry and materials science involves several steps to ensure the data is in a suitable format for model training. For example, Schilling-Wilhelmi et al.43 have demonstrated the importance of cleaning and segmenting text to focus on relevant sections. They used regular expression-based pipelines to remove unnecessary parts of the text, such as headers, footers, and references, which do not contain relevant information for NER tasks. Additionally, they employed semantic chunking to divide the text into meaningful segments, ensuring that each chunk contains relevant context. This approach helps in improving the efficiency and accuracy of the NER models by reducing noise and focusing on the most informative parts of the text.

Data annotation serves as the foundational step in constructing structured datasets for chemical information extraction. Given the inherent complexity of chemical texts—ranging from nuanced nomenclature of chemical entities to diverse reaction conditions—annotation methodologies must balance precision, scalability, and domain specificity. Current approaches predominantly fall into two paradigms: manual annotation by domain experts and automated annotation leveraging large language models (LLMs). This section delineates their workflows, comparative advantages, and empirical validations.

1. Manual Annotation Paradigm

Manual annotation relies on domain expertise to ensure high-quality labeled data, particularly critical in chemistry due to its specialized terminology and contextual dependencies. A representative implementation is exemplified by the ChEMU laboratory (CLEF 2020) during the construction of a chemical patent dataset [1].

Workflow:

1. **Initial Annotation**: Seed annotations are generated by mapping patent text to structured entries from authoritative databases (e.g., Reaxys), termed "silver-standard" annotations.
2. **Expert Review**: Two independent chemists validate entity spans (e.g., chemical substances, quantities) and augment missing entries. Discrepancies are flagged for arbitration.
3. **Consistency Arbitration**: Inter-annotator agreement (IAA) metrics (e.g., Cohen’ s κ) quantify discrepancies, with unresolved cases adjudicated by a third expert.
4. **Format Standardization**: Annotations are converted into structured formats (JSON/XML) using tools like BRAT, capturing entities (e.g., *Chemical*, *Condition*) and relationships (e.g., *ReactionStep*).

2. LLM-based Automated Annotation Paradigm

Recent advances in LLMs enable scalable annotation by synthesizing pre-trained knowledge with domain-specific tuning. As demonstrated in *Application of LLMs in Chemistry Reaction Data Extraction* [2], this paradigm enhances efficiency while maintaining robustness.

Workflow:

1. **Prompt Engineering**: Structured templates (e.g., JSON schemas) guide LLMs to extract entities (e.g., SMILES, molarity) and contextual attributes.
2. **Domain Fine-tuning**: Models like GPT-3.5 or LLAMA-2 are adapted using limited labeled data (100–500 samples) to improve chemical context recognition.
3. **Validation and Correction**: Outputs are filtered via computational checks (e.g., RDKit-based SMILES validation) and heuristic rules (e.g., discarding molarity outliers beyond ±20% expected ranges).

c. Model

Large Language Models (LLMs), or foundation models, show significant promise in tackling complex Natural Language Processing (NLP) tasks like Named Entity Recognition (NER) and relation extraction. The idea of "self-supervised learning" through transformer-based models like BERT, pre-trained on massive corpora of unlabeled text to learn contextual embeddings, is the dominant paradigm of information extraction today. These models can be fine-tuned on specific datasets for tasks like NER.

- **Fine-tuning**

    - **Method**: The process typically involves **pre-training a language model on a large amount of unlabeled text** using unsupervised objectives. The resulting encoder, such as a BERT-based encoder, generates token embeddings that are context-aware. These embeddings are then fed into a task-specific machine learning model, often a neural network (like a linear layer with softmax), which learns to predict the required labels, such as entity types. **Fine-tuning involves supervised learning** on specific, often labeled, datasets. For NER, this means feeding labeled inputs to the BERT model and using the output vector embedding for each word, along with its corresponding entity label, as input to a task-specific model. Examples of BERT-based models used or fine-tuned for materials science NLP tasks include BERT, SciBERT, BatteryBERT, MaterialsBERT, MatSciBERT, ChemBERTa, and BERT-PSIE. LLMs like GPT-3.5-Turbo, GPT-4, and GPT-4-Turbo have also been fine-tuned for NER and relation extraction tasks in materials science. Fine-tuning can involve adapting the model's behavior to a specific output format, such as a conversational style, before using another tool to transform the response into a structured format like JSON. For relation extraction fine-tuning, strategies can include varying the sorting of entity lists in the prompt or augmenting the dataset size with shuffled entity lists.

- **Prompt Engineering**
    - **Details**: Prompt engineering involves designing inputs to LLMs to guide their output. One approach is to use **prompt engineering with LLMs like ChatGPT** to structure text, generate summaries, and compile information. Another source mentions using LLMs with prompt-engineering and fine-tuning techniques.
    - **Chain of Thought (ACT)**: The sources do not explicitly mention "Chain of Thought (ACT)" as a prompting technique. However, one source describes an **"agent-based learning framework"** where an AI agent is equipped with tools, including a **"chain-of-verification (CoV)" tool**. This tool is used to autonomously re-evaluate the reasoning behind the agent's decision (e.g., predicting water stability) to ensure its logical connection to the property and reduce hallucinations. While not exactly the same as Chain of Thought prompting, it involves a reasoning/verification step.
- **Agent (LLM + Tools)**

    - **Toolkit**: An effective approach to address the limitations of standalone LLMs in intricate tasks is to augment them with **domain-specific toolkits**. This combination forms an AI agent. In one framework, the AI agent (named Eunomia in the case study) was equipped with various tools. A key tool is the **Doc Search tool**, which extracts relevant knowledge materials properties from text ranging from a single sentence to a full research paper. This tool works by embedding the paper and queries into numerical vectors and identifying the top 'k' passages within the document that mention or imply the property of interest. Another tool mentioned in conjunction with the agent is the **chain-of-verification (CoV)** tool.
    - **Other relevant aspects**: The integration of tools with LLMs aims to provide precise answers, addressing the inherent limitations of LLMs in specific domains and enhancing their overall performance and applicability. The agent framework was tasked with extracting information, such as identifying MOFs and predicting their properties. This approach was evaluated on tasks with increasing complexity, including NER, Relation Extraction, Template Filling, Argument Mining, and Entity Linking.

We provided a table(table 3) summarizing some model comparisons based on the provided sources in appendix.

d. Post-Processing

Post-processing steps are crucial after initial entity extraction to refine results, structure data, and handle complexities like co-references.

- **Validation**: The performance of NER models is typically **evaluated using metrics such as precision, recall, and F1 score**. These metrics compare the predicted entity tags to the ground truth labels. For materials science, evaluating extraction of intricate material expressions requires specific methods. A novel evaluation method involves **normalizing materials to their chemical formulas** before conducting a pairwise comparison of each element, referred to as "formula matching". This provides a more meaningful assessment for material names. Evaluation can also involve soft matching techniques to overlook minor discrepancies or strict matching. For datasets created through manual annotation by multiple experts, **inter-annotator agreement metrics** are used to assess annotation quality. Fact-based verifier metrics can also be used.

- **Output Form**: The goal is to transform unstructured text into a structured format. Extracted information can be compiled into a **highly structured format to form a materials database**. For the automated text mining pipeline for superalloys, the output was a 6-tuple relation: article DOI, alloy named entity, chemical element, content, property specifier, and property value. Generative models like LLMs can be trained to generate output in a **valid JSON (JavaScript Object Notation) format** as part of efforts to extract structured databases. Tasks can also be unified into **sequence-to-sequence formats** to facilitate the use of LLMs. Another output format mentioned is **ORD-formatted structured data**. For NER data specifically, the **CoNLL 2002 NER format** is used, which is a tab-separated text format using **Beginning–Inside–Outside (BIO) chunk tagging** to indicate the position of tokens within an entity.
- **Co-reference**: **Co-reference resolution focuses on finding all expressions that refer to the same entity** in the text. Examples include phrases like "these materials" or "each material" referring back to a previously mentioned specific material. Co-referencing material entity mentions across large spans of text and across figures and tables is a challenging problem. This task is critical for automated database development, especially when authors use abbreviations or simplified references after initially defining a compound. While LLMs have shown promise in other areas, one source notes that they have not been explored for intricate challenges like co-reference resolution. However, a fine-tuned LLM demonstrated the ability to recognize cross-referencing tokens.
- **Other relevant aspects**: A significant post-processing step is **named entity normalization**. This involves **identifying all the variations in the name for an entity across a large number of documents** and mapping them to a canonical form. For example, different names or acronyms for the same synthesis technique, like "chemical vapor deposition" and "CVD", should be recognized as synonyms. This normalization is achieved using methods like training a classifier to determine if two entities are synonyms, often using word embeddings and other features. Entity normalization greatly increases the number of relevant items identified in document querying.

e. Challenge

- **Ontology Expanding**: Ontologies are formal presentations of a domain, providing term meanings and hierarchical structures, and are useful for formalizing the semantics of entities. The ontology used in some work consists of important entity types but **misses other information about the material property record**, such as processing conditions, measurement methods, or conditions, which often influence property values. **Extending the ontology** to include this metadata would require explicitly labeling those new entities and training a new NER model, which is described as **time-intensive**. While supervised NER can extract entities like material names and properties using the current ontology, obtaining additional domain-specific information might require unsupervised methods like heuristic rules and regular expressions.
- **Fine-Grained**: NER tasks can operate at different levels of granularity. The MaterioMiner dataset, for instance, distinguishes between **coarse-granular NER (CG-NER)**, which recognizes high-level concepts from an ontology, and **fine-granular NER (FG-NER)**, which addresses the recognition of low-level concepts. This dataset is noted for its **eminently fine-granular annotation**, manually annotating 179 distinct classes. Recognizing and extracting this detailed, domain-specific information at a fine-grained level remains challenging.
- **Other relevant challenges**:
  - **Ambiguity and Complexity**: Scientific literature can contain **semantic ambiguity**, such as sentences with multiple mentions of compounds and properties, making correct association difficult for NLP. Complex operation sequences described in synthesis procedures, where steps might be omitted or branching occurs, are also challenging.
  - **Multi-modal Information**: Relevant material property information is **not confined to the text body but is also found in tables and figures**. Extracting data from these non-textual sources is challenging.
  - **Limited Data**: Many tasks, especially supervised ones, require a **large number of labeled samples or datasets**, which are often lacking or limited in materials science sub-domains. Extending methods to new domains requires labeling new datasets with tailored ontologies.
  - **Domain Specificity**: Generic NLP tools do not perform well in the materials science domain without modification due to the specialized vernacular, sentence structure, terminology, and chemical semantics. Materials names can have non-trivial variations. **Generalizing tools** across disparate sub-domains within materials science is challenging.
  - **Specific Entity Types**: Obtaining **SMILES strings for organic materials** is a bottleneck for generating structural fingerprints needed for property prediction models. Unlike polymers, organic molecules can often be converted to SMILES strings, but extracting them accurately is difficult.

- **Contextual Understanding**: NER relies on identifying the objects of semantic value. Materials-specific challenges vary by subdomain and include subtleties associated with the property, context, and reporting of measurements. **Non-local dependencies** (linked information across a document or modalities) are critical but difficult to capture with methods primarily relying on local dependencies.
- **Entity Linking**: Beyond recognition, **distinguishing between similarly named entities** (Entity Linking) is a challenge.

3. RE+Chemistry（陈子豪）

a. Domain-level tasks

In the fields of chemistry and materials science, Relationship Extraction (RE), as a core technology for accurately identifying and extracting predefined semantic relationships between chemical entities from vast amounts of unstructured text, plays a crucial role in constructing structured chemical knowledge bases and accelerating the paradigm shift towards data-driven scientific research. Through RE, researchers can uncover knowledge associations deeply embedded in literature, such as compound synthesis pathways, material structure-property relationships, and drug-target interaction mechanisms, thereby significantly advancing scientific discovery and technological innovation. RE tasks in the chemical domain are diverse, with their primary goal being to serve specific research and application needs.

The table below outlines the major RE task types in the chemical domain and their typical application scenarios:

Task Type	Description & Typical Examples
Chemical Reaction RE	Identifies complex interactions between reactants, products, catalysts, solvents, and reaction conditions (temperature, pressure, time, etc.). For example, extracting components and their roles, along with the "heating" condition, from "Ethyl acetate is hydrolyzed to sodium acetate and ethanol by heating in aqueous sodium hydroxide solution."
Material-Property RE	Extracts associations between specific materials and their physical properties (e.g., melting point, conductivity), chemical properties (e.g., stability, reactivity), or structural features (e.g., crystal phase, pore size). For example, identifying "MXene materials (Ti☒C☒T☒) exhibit great potential in energy storage due to their excellent conductivity and hydrophilicity."
Drug-Target-Disease RE	In medicinal chemistry and chemical biology, identifies interactions (e.g., inhibits, activates, binds to) between drug molecules, biological targets (proteins, enzymes, etc.), and related diseases. For example, extracting the drug, target, mode of action, and indication from "Osimertinib is a selective,

	irreversible EGFR tyrosine kinase inhibitor used to treat EGFR T790M mutation-positive non-small cell lung cancer."
Synthesis/Process RE	Identifies sequential or dependency relationships between precursors, intermediates, operational steps, equipment used, and final products in synthetic experiments or material preparation processes. For example, confirming "ZnO nanorods were synthesized by a hydrothermal method using zinc nitrate and hexamethylenetetramine as precursors, reacting at 90°C for 12 hours."

The effective execution of these RE tasks provides robust data support and analytical tools for automating the construction of chemical reaction networks, guiding new material design and screening, accelerating drug discovery processes, and empowering the development of chemical knowledge graphs.

b. Data annotation

High-quality annotated data is the cornerstone for training reliable supervised RE models and objectively evaluating their performance. In the chemical domain, the data annotation process is particularly complex and challenging. Firstly, it is necessary to precisely define the relationship schema to be extracted based on specific research or application needs. This includes the types of relationships and the entity categories involved. For instance, in studying extractant performance, one might need to define relationships like "`EXTRACTS(Extractant, MetalIon)`" and "`HAS_EFFICIENCY(ExtractionEvent, Value, Unit)`" to describe extraction efficiency. The rationality and completeness of the relationship schema directly impact the quality of subsequent annotation work and the utility of the model.

To ensure consistency and accuracy in annotation, detailed Annotation Guidelines are indispensable. These guidelines must provide clear definitions for each relationship type, typical positive examples, and easily confused negative examples, and clearly state the principles for handling ambiguous or complex situations, such as how to differentiate between catalysts and reaction promoters, or how to deal with implicitly stated causal relationships in literature. The specialized nature of chemical text requires annotators to possess relevant chemical background knowledge and to improve annotation quality through systematic training and continuous feedback calibration. Metrics such as Cohen's Kappa or Fleiss' Kappa are commonly used to quantify Inter-Annotator Agreement (IAA), serving as an important basis for assessing data reliability. Professional text annotation tools like BRAT, Doccano, or INCEpTION significantly enhance annotation efficiency and ensure uniformity in annotation format by providing graphical annotation interfaces and standardized data export functions. However, the complex and diverse expressions of relationships in chemical literature, such as long-range dependencies, coordinate structures, implicit information, and negative expressions, make annotation work time-consuming, labor-intensive, and costly. Concurrently, the rapid development of knowledge in the chemical domain, with the continuous emergence of new compounds, reactions, and materials, poses ongoing challenges to the dynamic adaptation and updating capabilities of annotation schemas.

c. Model

The development of models for relationship extraction in chemistry has progressed from traditional machine learning methods to deep learning approaches, and further into a new phase empowered by large-scale Pre-trained Language Models (PLMs) and Large Language Models (LLMs). Initially, rule-based methods relied on domain experts to manually construct patterns and rules; while achieving high precision in specific scenarios, their generalization ability and recall were limited. Subsequently, feature-based statistical learning methods, such as Support Vector Machines (SVM) and Maximum Entropy models (MaxEnt), performed relation classification by designing features related to vocabulary, syntax (e.g., dependency paths), and entities, but the feature engineering process itself was time-consuming and highly dependent on expert experience.

The advent of deep learning methods, with their powerful ability to automatically learn features, has made them mainstream in the RE field. Convolutional Neural Networks (CNNs) excel at capturing local contextual features in text; Recurrent Neural Networks (RNNs) and their variants (such as LSTM, GRU) are better suited for processing sequential information and long-range dependencies. The introduction of the Attention Mechanism has significantly enhanced the performance of architectures like RNNs and Transformers by allowing models to focus on key parts of a sentence relevant to a specific relationship. Furthermore, when text can be converted into graph structures (e.g., dependency parse trees) or when external knowledge needs to be integrated, Graph Neural Networks (GNNs) have demonstrated advantages in learning entity and relationship representations.

In recent years, Pre-trained Language Models (PLMs), particularly BERT-derived models pre-trained on scientific literature or specifically for the chemical domain (e.g., SciBERT, ChemBERT), have vastly improved extraction performance. They achieve this by learning rich language representations from massive text corpora and then fine-tuning on downstream RE tasks, showing particular strength in low-sample scenarios and domain adaptability. A common paradigm involves feeding the representations of entity pairs in a candidate relation into a classification layer. More recently, prompt-based learning approaches, which reformulate RE tasks into formats similar to pre-training tasks (e.g., cloze tests) using templates, have effectively unlocked the potential of PLMs in few-shot or even zero-shot scenarios.

Currently, Large Language Models (LLMs), such as the GPT series, are bringing new breakthroughs to the RE field with their exceptional contextual understanding, reasoning, and generation capabilities. Through carefully designed instructions or by leveraging their in-context learning abilities, LLMs can perform relationship extraction directly in zero-shot or few-shot settings, and can even handle more complex, open-ended relationship types. This, to some extent, alleviates the dependency on large-scale annotated data. However, ensuring the precision and controllability of LLMs in specific chemical sub-domains, as well as effectively mitigating their "hallucination" problem, remain key research focuses and challenges.

d. Post-Processing

After the model outputs initial extraction results, implementing effective post-processing steps is crucial for enhancing the quality, consistency, and usability of the final output. This stage typically involves rule-based filtering and validation based on chemical domain knowledge, such as using atom conservation, valency rules, or constraints of known reaction types to eliminate obviously erroneous or illogical relationship triples. For example, if an extracted reaction shows an element present in the reactants but completely absent in the products, this extraction result is likely incorrect.

Many models also provide a confidence score for extracted relationships. By setting an appropriate threshold, a trade-off can be made between precision and recall, removing low-confidence results that are more likely to be erroneous. For relationship instances extracted from different text segments but describing the same fact, relationship deduplication and fusion operations are necessary to ensure knowledge non-redundancy and completeness. For instance, information about the same reaction step extracted from both the abstract and the main body of a paper should be merged.

Furthermore, aligning and enriching the model's extracted relationships with existing authoritative chemical knowledge bases (e.g., ChEMBL, PubChem, Reaxys) can not only validate the accuracy of the extractions but also supplement and enrich the new results with structured information already present in these KBs, such as completing standardized compound names or CAS numbers. For critical application scenarios or extraction results where the model exhibits high uncertainty, introducing a human review and iterative refinement mechanism is essential. Domain experts review and correct these results, and the corrected high-quality data is fed back to the model for incremental learning or fine-tuning, thus forming a continuous improvement loop that progressively enhances the model's performance and reliability.

e. Challenge

Despite significant advancements in RE technology for chemical information extraction, its practical application still faces multifaceted challenges, which collectively constitute key directions for future research.

Challenge Category	Specific Description
Complexity and Ambiguity of Chemical Text	Chemical literature is replete with diverse compound nomenclatures (systematic names, common names, trade names, CAS numbers), numerous abbreviations, and complex descriptions of chemical reactions and material preparation processes. Long sentences, intricate syntactic structures, and implicitly stated, non-direct relationships between entities (e.g., causal relationships inferred through multiple steps) are prevalent. Furthermore, accurately identifying and processing negative expressions (e.g., "no catalytic activity was observed") and uncertainty descriptions (e.g., "it is speculated that a complex may be formed") poses significant demands on the precision of RE.
Data Sparsity and Annotation Cost	High-quality, large-scale annotated datasets for specific chemical sub-domains or fine-grained relationships remain scarce. The specialized nature of chemical text requires annotators to possess deep domain knowledge, making the annotation process not only time-consuming and labor-intensive but also difficult to rapidly scale to new research areas or relationship types due to high labor costs.
Long-range and Complex Dependencies	Relevant entities or pieces of information required to form a complete relationship may be distributed across different parts of a sentence or even span multiple sentences or paragraphs. Effectively capturing and understanding these long-range, non-local contextual dependencies poses a severe test for a model's sequential understanding and information integration capabilities.
Fine-grained Relation Distinction & Novel Relation Discovery	Accurately distinguishing between semantically similar but substantially different fine-grained relationship types (e.g., differentiating the subtle nuances between "catalyzes," "promotes," and "enhances") is a major difficulty. Concurrently, most existing RE methods rely on predefined relationship schemas, making it challenging to automatically discover novel, undefined relationships in literature or to identify entirely new categories of chemical entities, thereby limiting the boundaries of knowledge discovery.
Multimodal Information Fusion	Chemical literature typically contains information in multiple modalities, including text, tables, chemical structure diagrams, reaction flowcharts, and spectra (e.g., NMR, IR, MS). Effectively fusing this multimodal information from diverse sources and structures to enhance the accuracy, coverage, and robustness of RE is a highly challenging frontier research direction.
Model Interpretability, Robustness, and Trustworthiness	Deep learning models, especially large language models, often have decision-making processes that lack transparency and are considered "black boxes," making it difficult to explain why a particular relationship was extracted or why an error occurred. Simultaneously, the robustness of models to minor perturbations in input text (such as spelling errors, synonym replacements, sentence paraphrasing), as well as the veracity (avoiding "hallucinations") and

	trustworthiness of their generated content, are critical issues awaiting resolution for practical applications.

Overcoming these challenges will further propel chemical information extraction technology towards being more intelligent, precise, practical, and trustworthy, thereby better serving chemical research and industrial applications.

4. Joint NER+RE+Chemistry（杨婷然）

a. Domain-level tasks

In the field of chemistry, the joint application of Named Entity Recognition (NER) and Relation Extraction (RE) has become a core method for information extraction, enabling effective extraction of chemical knowledge from a vast amount of literature. These techniques can not only identify key entities such as chemical substances, reaction conditions, and experimental results, but also extract the complex relationships between these entities. This allows unstructured text to be transformed into structured information, significantly advancing research and discoveries in fields like materials science, pharmaceutical chemistry, and catalysis. For example, in chemical reactions, NER can identify key entities such as reactants, products, and catalysts, while RE can extract the relationships between these entities, such as the transformation between reactants and products and the catalytic role of catalysts. In pharmaceutical chemistry, NER can extract the names, structural information, and pharmacological properties of drug molecules, while RE can reveal the relationship between molecular structures and their biological activities. In materials science, NER can identify different types of materials, synthesis conditions, and physical-chemical properties, while RE can extract the relationships between materials and their performance, helping researchers better understand material behaviors. In battery research, NER can identify key components such as anode materials, cathode materials, and electrolytes, while RE can extract the relationships between these components and battery performance, such as capacity and cycle stability. In catalytic reactions, NER can identify catalysts, reactants, and products, while RE can reveal catalytic roles and the relationships between catalysts and reactants.

This combined NER and RE approach not only efficiently extracts key information from literature but also establishes hierarchical relationships between entities, making chemical knowledge more systematic and suitable for subsequent data analysis and model building. With the advent of large language models (LLMs) like GPT and LLaMA, the performance of these joint tasks has become increasingly impressive, handling more complex multi-level relationships and enabling more precise information extraction. Unlike traditional NER and RE methods, this approach does not require the enumeration of all possible n-tuples of relations. Instead, it fine-tunes large language models to flexibly handle complex hierarchical relations and extract them accurately, providing stronger adaptability. Furthermore, this approach allows users to define the output structure, generating structured representations of knowledge that are easy to process.

Domain-level Task	Targeted Entities	Relationships
Chemical Reaction Extraction	Reactants, Products, Catalysts, Solvents	Transformation between reactants and products, catalytic effect of catalysts on reactants
Pharmaceutical Chemistry & Molecular Property Extraction	Molecules, Molecular Structures, Pharmacological Properties, Experimental Conditions	Relationship between molecules and their properties, e.g., "anticancer activity," "optimal concentration"
Materials Science: Properties & Processing Conditions Extraction	Materials, Synthesis Conditions, Physical Properties, Characterization Methods	Relationship between synthesis conditions and physical properties, e.g., "temperature and hardness"
Battery Research: Electrode Materials & Performance Extraction	Anode Materials, Cathode Materials, Electrolytes, Battery Performance	Relationship between electrode materials and battery performance, e.g., "capacity and current density relationship"
Catalytic Reactions: Catalyst & Reactivity Extraction	Catalysts, Reactants, Products, Experimental Conditions	Relationship between catalysts and reactants, e.g., "catalyst and product relation"
Virology & Epidemiology: R_0 Estimation Extraction	Disease Name, Location, Date, R_0 Value, Confidence Interval (%CI), Estimation Method	Structured representation of R_0 studies through property-value pairs; relation between disease and R_0 estimate across different contexts

b. Data annotation

In the field of joint Named Entity Recognition (NER), Relation Extraction (RE), and Chemistry, data annotation is a critical yet complex process that directly determines the quality of model training and evaluation. To ensure practical and high-quality annotations, researchers often begin by selecting a curated subset of documents from large scientific corpora. These documents are then manually annotated to create gold-standard datasets, typically structured in both plain-text and JSON formats to accommodate downstream applications. The annotation process involves identifying chemical entities such as materials, reagents, and experimental conditions, and establishing relationships among them (e.g., catalyst–reactant, temperature–reaction rate). Despite the labor-intensive nature of the task, much of the information can be directly extracted from the text, making the work tedious but not highly ambiguous. However, variability in the linguistic expression of chemical information presents a unique challenge; for instance, “Pd-intercalated silica” and “silica intercalated with Pd ions” convey the same concept in different forms. To account for such variations, manual evaluation scores are often introduced to complement strict automated metrics like precision and recall, ensuring that semantic correctness is preserved even when string matches differ. Furthermore, human-in-the-loop workflows are widely adopted to accelerate annotation by using model predictions as initial labels that are later refined by human annotators. This iterative strategy reduces effort and enables efficient expansion of training sets. Lastly, transparency and reproducibility are emphasized by open-sourcing annotated datasets, preprocessing scripts, and model weights. Together, these practices contribute to building robust, semantically accurate, and practically usable datasets for chemical information extraction in scientific texts.

c. Model

In the domain of structured information extraction for chemistry and materials science, some studies focus on jointly modeling Named Entity Recognition (NER) and Relation Extraction (RE), leveraging Large Language Models (LLMs) for efficient implementation. One approach involves instruction tuning, using encoder-decoder LLMs such as FLAN-T5, with a representative example being ORKG-FLAN-T5R0. This method formulates the task of extracting structured triples from scholarly text as a sequence-to-sequence problem, where the input includes carefully crafted natural language instructions and the output consists of the extracted structured information. It explores the impact of different types of instructions on model performance, including single-instruction training, multi-instruction joint training, and selecting the best-performing instruction based on validation results. Although the model does not rely on extremely large parameter sizes (e.g., 11B), it achieves robust performance on complex extraction tasks with only 780M parameters. To support various downstream tasks, ORKG-FLAN-T5R0 offers both plain-text and JSON outputs, and evaluates its performance using exact-match and partial-match Precision/Recall/F1 metrics under fine-tuning settings, as well as ROUGE and other summarization-style metrics under zero-shot settings.

In addition, a schema-driven fine-tuning strategy can be employed to build task frameworks more closely aligned with the semantic structures of materials science. Representative works utilize decoder-only models like GPT-3 or Llama-2 and perform information extraction through prompt-based completion. This approach places particular emphasis on the design of schemas, framing the task as generating structured entities and their attributes or relationships from a single prompt. For instance, in the task of material doping, the schema explicitly distinguishes elements such as Host, Dopant, Result, and Modifier, and supports many-to-many mappings between entities. In domain-specific applications like Metal–Organic Frameworks (MOF), the model identifies semantic links between material names and their application contexts. Furthermore, researchers have developed general-purpose schemas for broader materials corpora, covering elements like formulas, applications, and crystal structures. During data construction, a human-in-the-loop approach is often adopted, where experts manually annotate a subset of the data, and fine-tuned models generate candidate outputs for human correction—enhancing data quality and coverage. The final outputs are typically in JSON or natural-language-like structured formats, facilitating subsequent use in knowledge graph construction or database integration.

d. Post-Processing

In the domain of joint Named Entity Recognition (NER), Relation Extraction (RE), and Chemistry, post-processing plays a vital role in transforming raw model outputs into clean, structured, and semantically accurate representations. This stage typically involves several tasks such as entity normalization, error correction, format validation, and hallucination mitigation. For instance, chemical entities extracted directly from the text often contain extraneous whitespace or inconsistent naming; a post-processing step might correct "Li Co O2" to the standardized "$LiCoO_2$," or align "PdO functionalized with platinum" to a structured JSON format like `{formula: "PdO", description: ["Pt-functionalized"]}`. Furthermore, post-processing ensures that the output adheres to a predefined schema (e.g., JSON), which is crucial for downstream applications such as knowledge graph construction or automated literature analysis. In some cases, model hallucinations—where the LLM fabricates information not present in the source—must be filtered out, especially when the generated content is chemically plausible but unverifiable in context. Notably, recent methods like LLM-NERRE have begun to embed these normalization and correction procedures directly into the training data, enabling the model to learn how entities should appear and be structured, thereby reducing the need for extensive post-processing. This approach not only improves consistency but also significantly accelerates the annotation and model deployment pipeline. Nonetheless, post-processing remains an essential safeguard, particularly in scientific domains where precision and format compliance are critical.

e. Challenge

In the field of chemistry, joint Named Entity Recognition (NER) and Relation Extraction (RE) tasks face numerous challenges and limitations, primarily due to the complexity, multimodality, and unique expression styles of chemical data. First, chemical literature involves diverse types of entities with inconsistent naming conventions, including a mix of technical terms, abbreviations, systematic nomenclature, and common names, which poses significant difficulties for NER. Moreover, relationships between chemical entities often rely on complex experimental setups, data tables, image annotations, or cross-references within the context, making it difficult for a single model to capture all relevant information accurately. Most existing joint NER+RE methods are primarily text-based, yet in chemical documents, key information is frequently embedded in figures, formulas, and structural diagrams (e.g., molecular structures, CIF files), which current models struggle to handle robustly and generalize across. Although vision-language models (VLMs) offer a pathway to extract data from both text and visual sources, their adaptability to the chemical domain remains limited—especially when it comes to mapping structural data to semantic information. Compounding the issue, chemical literature often relies on cross-document or external knowledge base references, where the meaning of an experiment or entity may depend on definitions or results in other papers. Current joint extraction models are typically constrained to single-document contexts and cannot effectively reconstruct knowledge graphs across documents. Additionally, scientific publishing is biased toward positive results, with negative or failed experiments rarely reported. This leads to training data that lacks full coverage of real-world scientific processes, hindering the development of comprehensive extraction systems. The absence of high-quality, multimodal, and cross-document benchmarks for joint NER+RE further complicates progress—the tasks are often overly simplified and fail to reflect the actual extraction needs in real-world chemistry. Finally, while structured scientific publishing (e.g., semantic publishing) offers promise, it is still in its early stages. A lack of standardized formats, tools, and broad community adoption limits the development of large-scale, high-quality chemical knowledge graphs. Thus, joint NER+RE in the chemistry domain not only encounters modeling and data-level challenges but also requires coordinated advancements in methodology, tooling, and community consensus.

5. EE+Chemistry（薛柔）

a. Domain-level tasks

In this review, EE stands for Event Extraction. It refers to the task of identifying and extracting information related to specific events from unstructured text. The goal of EE is to detect the event itself and then capture the relevant attributes associated with the event.

Domain-level Task	Targeted Entities
General Materials Information Extraction	Synthesis processes and parameters, Process routes
Synthesis Procedure Extraction (General Inorganic Materials)	Chemical synthesis procedures, Targets, Precursors, Operations, Conditions, Reactions, Components, Reaction conditions, Flowcharts of possible synthesis procedures
Synthesis and Processing Extraction	Synthesis and processing actions, Parameters

(Continuous Events, Linked in Sequence)	
Experimental Procedures Extraction	Action sequences
High-fidelity Chemical Data Extraction	Reaction-related information

b. Data annotation

Data annotation for these extraction tasks often involves manual labeling, although challenges exist, particularly for complex descriptions like alloy processing routines or when handling token/chunk-level actions.

Semi-supervised methods have been proposed and used to expand labeled data or automatically generate corpora to improve extraction precision. For example, a semi-supervised Snorkel framework was used for the materials domain via automatically generated corpus.

The limited size of hand-labeled datasets, especially for niche areas like alloys, presents challenges for training supervised deep learning methods for tasks like named entity recognition or relation extraction related to processes.

For LLM-based approaches, prompt engineering can be used to direct the models for extraction, which can differ from conventional NLP pipelines that rely on explicitly annotated data for training. Some LLM approaches leverage "human-in-the-loop" annotation processes.

c. Model

Traditional NLP pipelines for automatic data extraction have been developed, often relying on rule-based approaches, machine learning (ML) methods, or a hybrid combination.

Specific ML models like neural networks and parse-based methods have been used to extract synthesis parameters. BiLSTM-CRF models were used for named entity recognition related to materials science entities (which would be part of event extraction).

Semi-supervised machine-learning methods like latent Dirichlet allocation and Markov chains have been applied to classify synthesis procedures and reconstruct flowcharts.

Transformer-based language models, including BERT-based models (like BatteryBERT) and fine-tuned versions, have been used, offering improved context-aware interpretation.

Large Language Models (LLMs), such as GPTs and Llama models, are increasingly used, often via prompt engineering or fine-tuning with domain-specific data.

Multimodal models (VLMs) are emerging to extract information from non-textual data modalities critical for synthesis procedures, such as reaction schemes and figures.

AI agents are being explored to autonomously perform complex research tasks, including information extraction and experimental execution, by integrating LLMs with retrieval tools and potentially controlling laboratory equipment. An autonomous AI agent framework has been developed specifically for chemical literature data mining to extract reaction-related information.

d. Post-Processing

Post-processing often involves interdependency resolution among extracted entities.

Data normalization is crucial, especially for chemical names (using tools like PubChem) and units (using tools like pint or unyt) to ensure consistency before analysis or comparison.

Validation using chemical knowledge and understanding is a significant advantage in this domain. This involves applying "sanity checks" based on domain rules and links between data entries. Examples include using cheminformatics tools to validate consistency of extracted molecular properties or checking the conservation of atoms in extracted reaction equations. This validation can also serve as an early evaluation loop.

Using another LLM to check for factual inconsistencies or hallucinations in the extracted data is also possible.

The extracted structured data can be used to populate databases or create structured summaries and knowledge graphs.

e. Challenge

In the process of EE in the field of chemistry, there are lots of challenges:

Handling the description of synthesis and processing routes as continuous events with actions linked in sequence, which exhibit diverse types, flexible expressions, and varied conditions/parameters. The diversity of topics and reporting formats in chemical and materials research presents a challenge for traditional methods hand-tuned for specific cases. While LLMs handle variability better, they can struggle with domain-specific polysemy (ambiguous terms like "yield"). Handling diverse data modalities beyond text, such as reaction schemes, tables, figures, crystal structures, and spectra, which contain critical information related to procedures and outcomes is also a difficult thing.

Distinguishing between token-level and chunk-level action entities and the variability in how actions are described based on their position in the process. The descriptions of events are frequently intertwined with discussions on experimental phenomena and intermediate products, making extraction difficult.

Ensuring models provide accurate and reliable predictions or extractions, as they often lack the specificity and domain expertise required for intricate tasks. Establishing ethical and validation frameworks for AI-generated content and hypotheses derived from extracted data, including auditing AI-proposed reactions.

The lack of standardized benchmarks for structured data extraction makes systematic evaluation challenging.

6. Application（杨婷然）

In the field of chemistry, the application of Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), and Joint NER+RE technologies is significantly driving the automation and informatization of chemical research. The primary task of NER is to identify and extract chemical entities from literature, such as compound names, chemical formulas, experimental conditions, and more. For example, NER can automatically recognize chemical substances like "sodium chloride (NaCl)" or "benzene (C_6H_6)," as well as experimental conditions like temperature and time involved in reactions. This not only helps researchers quickly organize key information from the literature but also supports database construction and literature retrieval. RE focuses on identifying various relationships between chemical entities, particularly in chemical reactions. Through RE, researchers can extract relationships such as "sodium chloride reacting with water to form hydrogen chloride," thereby helping to establish reaction networks and reveal the rules of chemical reactions. In addition, RE can be used in pharmaceutical chemistry to identify the relationships between molecules and biological targets, aiding in drug discovery and molecular design. EE, on the other hand, specializes in identifying key events in chemical experiments or processes, such as changes in reaction conditions or experimental steps. EE can automatically extract information about reaction temperature, pressure, reaction time, and other factors from the literature, helping to accelerate experimental design and optimization. Additionally, EE can assist in extracting biological events in drug research, such as "compound X inhibits enzyme Y activity," providing crucial pharmacological information. The Joint NER+RE technology combines these two tasks, allowing for more comprehensive and efficient extraction of information from chemical literature by simultaneously recognizing chemical entities and extracting relationships. This method allows researchers to quickly identify chemical entities and extract their complex relationships, providing essential support for reaction pathway prediction, new material design, and the construction of chemical knowledge graphs. Overall, the application of these technologies greatly improves the efficiency and accuracy of information extraction, providing strong technical support for research, innovation, and discovery in the field of chemistry. With the rapid development of big data and artificial intelligence, NER, RE, EE, and Joint NER+RE will further drive the automation of chemical research, having a profound impact on the discovery of new compounds, drug design, and materials science research.

7. Future（薛柔）

a. Possible improvements to information extraction

Future advancements will enable LLMs to achieve greater success by enhancing fundamental capabilities like numerical reasoning, quantitative predictions, and structural interpretations. Systematically enhancing numerical capabilities requires focusing on dataset construction, model architecture, task planning, training optimization, and tool integration. Establishing quantitative relationships between composition, processing, and properties from text is a challenge that future strategies, such as integrating materials language encoders with property prediction networks or using AI agents with computational tools, aim to address. Scientific reasoning needs improvement to mitigate the generation of inaccurate or hallucinated information, which can be addressed through strategies like Retrieval-Augmented Generation (RAG) and leveraging reinforcement learning (RL).

LLMs and Vision Language Models (VLMs) are expected to become more robust in handling diverse data modalities beyond just text, such as tables, figures, crystal structures, reaction schemes, and spectra. This includes addressing the complexity of diverse modalities contained within different data structures. A key frontier is tackling cross-document linking, where current methods primarily focus on single documents. Future approaches, potentially using multiple agents and RAG, will aim to handle complex relationships and information contained across multiple documents and knowledge sources, understanding references and their context. Developing multimodal LLMs that holistically integrate text, molecular graphs, spectra, and experimental data is a critical avenue.

LLMs are currently seen as tools for expediting exploration. However, their role could evolve to become sources of new insights and discoveries as they become more sophisticated and versatile. The full potential of LLMs remains untapped. Future research aims to transition LLMs from assistive tools to autonomous discovery engines capable of generating testable hypotheses. These systems could bridge disciplinary silos and accelerate the translation of knowledge into real-world innovations.

Ultimately, the hope is that the evolution of NLP and LLMs will not only streamline the materials design process but also foster innovative breakthroughs to significantly reduce the time and costs of materials discovery. The future lies in amplifying human ingenuity through "linguistically intelligent systems that speak the language of molecules".

b. Future scientific work in the fields of chemistry with IE

The vast majority of materials knowledge is published as scientific literature in unstructured text format. Information extraction is critical for converting this unstructured data into organized, structured formats, such as databases, which are crucial for innovative and systematic materials design.

This automated data construction is a necessity, addressing the severe limitation and time-consuming nature of manual data collection from the overwhelming volume of literature.

Specifically, information extraction from text, including compounds, compositions, properties, synthesis processes, parameters, and process routes, will enable:

- Materials Discovery and Design: Creating databases for data-driven materials design, assisting in new materials discovery, identifying candidates, and enabling the generation of new molecular structures and compounds.
- Property Prediction: Extracting data and encoding it (e.g., as word embeddings) can establish relationships with properties and be used directly for predicting materials properties.
- Composition and Process Optimization: Extracting details on compositions and synthesis routes enables efforts to identify optimal compositions and suggest processing conditions.
- Knowledge Synthesis and Question Answering: Extracting and processing information from vast texts allows LLMs to answer precise questions about chemical concepts and summarize extensive research reports.

Besides, by integrating extraction with search and machine learning agents, systems could autonomously find data and train models to answer research queries. Ultimately, enhanced information extraction is a key component in enabling LLMs to transition into autonomous discovery engines capable of generating testable hypotheses. This promises to streamline the materials design process, foster innovative breakthroughs, and significantly reduce the time and costs of materials discovery.

In summary, information extraction is fundamental because it unlocks the knowledge hidden in unstructured scientific literature. In the future, powered by sophisticated NLP and LLMs that can handle diverse data types, contexts, and integrate with other tools, it will not just provide data but directly contribute to discovery, prediction, design, optimization, and the eventual realization of autonomous research workflows in chemistry and materials science.

附录

Table 2:Overview of some data sources relevant for structured data extraction from scientific text, including published articles in open-access archives and data dumps

Table 3:

Method	Model	Normal Form	Model Size	Dataset Name	Dataset Size	Task

A document-level information extraction pipeline for layered cathode materials for sodium-ion batteries	BERT-based models (SciBERT, BatteryBERT), ChemDataExtractor (BERT, CRF, rules, dictionaries), Heuristic rules	Fine-tuning, Mixture of model and rule processing, Heuristic rules	BERT-based, BatteryOnly BERT, BatteryBERT	Corpus of layered cathode materials for SIBs, Manually annotated abstracts, Categorized paragraphs	**1747 documents**, 5265 property records, 1140 train/286 test abstracts, 50 documents for manual check	IE, Text classificati n, NER, Relation Extraction, Extracting relationshi s <chemica name, properties, physical parameters, Synthesis/ Test condition extraction, Cycling/ Rate performan e property relation, Binary document classificati n, MLTC, Abbreviati n definitio
A general-purpose material					Large polymer corpora, **750**	Extracting material property info from abstracts,

property data extraction pipeline from large polymer corpora using natural language processing	MaterialsBERT, BERT, PubMedBERT, ChemBERT, MatBERT, Heuristic rules	Fine-tuning, IO tagging scheme, Heuristic rules, Manual annotation	BERT-base architecture	Corpus of papers (abstracts), PolymerAbstracts	**annotated abstracts** (85% train, 5% val, 10% test), MaterialsBERT pre-trained on 14M abstracts + full text / 2M articles	NER (8 entity types), Relation Extraction, Extracting material property records, Extracting specific property data
A rule-free workflow for the automated generation of databases from scientific literature	BERT-PSIE (fine-tuned BERT models), Hybrid ChemDataExtractor	Fine-tuning, Rule-free workflow, Hybrid approach, Sentence-level processing	N/A	Test set, Manually curated databases, Corpus of scientific papers, Dataset for fine-tuning relation classification	Test set size N/A, Approx. 77,000 papers downloaded (band-gap), Approx. 126,000 sentences (band-gap), Final database 2,090 unique records (band-gap), Fine-tuning relation dataset 200 sentences	Automated database generation, Extracting structured data, Minin TC/Band-gap, Binary related IE, Sentence classificati n, NER, Relation Extraction, Relation classificati n

Agent-based Learning of Materials Datasets from Scientific Literature	AI Agent (Eunomia), LLM-NERRE (comparison), OpenAI text-ada-002 embeddings, Cohere embed-english-v3.0 embeddings	Agent-based learning, Zero-shot, Chain-of-verification (CoV), Doc Search, Exact word matching (evaluation)	N/A	Hand-labeled dataset based on MOF papers, Dataset for water stability, Case Study 1/2 datasets	**101 materials research papers**, **371 MOFs**, Case Study 1/2/3 sizes N/A for full datasets, Support counts vary per entity	Informatio extraction, Identifying MOFs/ properties (water stability), NER, Relation Extraction (host-dopants, MOF formula-guest species), Co reference resolution, Argument mining, Entity linking, Template Filling, Extracting relevant context
Annotating and Extracting Synthesis Process of	Deep learning sequence tagger, Rule-based	Deep learning sequence tagging,	N/A	SynthASSBs corpus, Experimenta	Experimental sections of **243 papers**,	Annotating Extracting Synthesis Process of ASSBs, Extracting synthesis processes, Defining flow graphs

All-Solid-State Batteries from Scientific Literature	based relation extractor, Sequence tagger with Mat-ELMo	Rule-based relation extraction, Heuristic rules	N/A	l sections of papers, Development dataset	Development dataset size N/A	Entity detection, Relation extraction, Extracting relation between PROPERTY/ MATERIAL/ OPERATION
Automated extraction of chemical synthesis actions from experimental procedures	Combined rule-based model, Pretrained/ Refined translation models, Model without pretraining	Rule-based, Translation model	N/A	Annotation test set, Dataset used by Vaucher et al.	Annotation test set size N/A	Automated extraction of chemica synthesis actions, Translating procedures to action sequences, Extracting action sequences/ properties
					14425 journal articles,	Extracting

Automated pipeline for superalloy data by text mining	Rule-based NER, BiLSTM-CRF (comparison), ChemDataExtractor (comparison), Heuristic distance-based RE, Table parse/RE algorithms	Rule-based NER, Heuristic algorithm, Table parsing/classification, Text classification, Requires no labeled corpus	N/A	Corpus of journal articles, Small corpus, Superalloy documents, Sentences for RE evaluation, Articles for sentence classification, Table cells for manual inspection	Small corpus size N/A, ~**14000 articles**, 329 sentences (RE eval), 30 articles (~3000 sentences) (sentence class), 4593 table cells (manual inspection), 9158 tables extracted, 5327 composition tables, 114 solvus tables, 743 complete records	superalloy data, Text mining, NE (alloy, property specifier, value), Relation Extraction (text/table, 6-tuple), Interdepen ency resolution, Table parsing/ classificati n, Text classificati n, Extractin chemical compositio /properties
ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis	ChatGPT	Prompt-engineering (ChemPrompt), Tabulation as output, N/A for missing data	N/A	Experimental section paragraphs from MOF synthesis papers	Size not explicitly stated	Converting paragraph to table, Extracting 11 MOF synthesis parameters Summarizi g synthesis conditions, Determinin quantities/ volumes

ChemNLP	ChemDataExtractor, JARVIS-Tools, Named entity recognition model, BERT-based, SciBERT	Pre-training, Fine-tuning, Use of output vector embeddings	BERT-based, SciBERT	arXiv:cond-mat dataset, Massive corpora (unlabeled), Specific datasets (fine-tuning)	Massive corpora, arXiv:cond-mat size N/A	Informatio extraction, Text classificati n, Entity recognitio (NER), Question answering, Relation extraction, Recognizin organic/ inorganic entities, Abstract summariza on (implied
Extracting structured data from organic synthesis procedures using a fine-tuned large language model	Fine-tuned LLM, Closed-source LLMs, Transformer-based model (Vaucher et al.), SynthReader (rule-based)	Fine-tuning, In-prompt data schema, Rule-based	N/A	Organic synthesis procedures, ChEMU evaluation lab corpus	Size N/A	Extracting structured data, Converting procedures to action sequences, Translating to cDL, NER RE (reaction/ workup steps), Extracting parameters objects

Fine-tuning large language models for chemical	Large Language Models (LLMs), GPT-3.5-turbo, GPT-4.0, Llama 2, T5 (comparison), BART (comparison), BERT-like models	Fine-tuning, Prompt engineering, Sequence-to-sequence, LoRA	7b, 13b (Llama-2), GPT-3.5-turbo, GPT-4.0	Datasets for 5 chemical text mining tasks, Paragraph2 MOFInfo dataset, Re-annotated Paragraph2 MOFInfo, USPTO reaction dataset (mentioned)	Dataset details in Table S1, Paragraph2 MOFInfo: 329 train, 329 test samples. USPTO size N/A.	Chemical text mining tasks, Labeling reaction roles, Extracting MOF synthesis informatio (11 parameter - Paragraph2 MOFInfo), Extracting NMR data, Converting procedures to action sequences
MATERIOMINER	Pre-trained models (for fine-tuning NER), BERT-based, Seq2seq BART (REBEL), YOLOv5, TableModel	Fine-tuning, Token classification (NER), BIO tagging, Direct triple extraction	BERT-based, BART, YOLOv5, TableModel	MaterioMiner dataset, Literature corpus (materials mechanics), Materials science datasets (comparison). CoNLL	**4 publications**, **2191 entities**, 12,155 tokens, 27% annotated, 179 distinct classes (FG-NER). Comparison datasets: 45 pubs (288k tokens, 2% anno), 750 abstracts (24k tokens,	NER (coars fine-granular), Token classification, Relation Extraction (RE), Entity linking (no NEL), Fine-tuning for triple extraction, Causal relationshi

	(LSTM, attention)			2002 NER, Webanno TSV	14% anno), 800 abstracts (111k tokens, 20% anno), 305 pubs (66k tokens, 22% anno).	extraction (conceptua, PDF text/ image/tabl extraction, Classifying PDF cells
Mining experimental data from Materials Science literature with Large Language Models	LLMs (GPT-3.5-Turbo, GPT-4, GPT-4-Turbo), BERT, Rule-based algorithm (baseline), Sentence BERT	Zero-shot, Few-shot, Fine-tuning, Strict matching (RE eval), Formula matching (evaluation), JSON output required	GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, BERT, Sentence BERT	SuperMat, MeasEval, grobid-quantities dataset	SuperMat RE fine-tuning: 344/148 (base/doc order), 695/299 (augmented). SuperMat NER fine-tuning: 1639/703. grobid-quantities NER fine-tuning: 485/208. Evaluation dataset size N/A, Support counts provided.	Informatio extraction, NER (materials, properties) Relation Extraction (materials-properties) Evaluating IE on materials entities, Text classificatio n
				Corpus of		Large-scale IE

Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature	Neural network (NER), Supervised ML (Normalization), Linear classifier (Document selection), ChemDataExtractor	Supervised ML, Trained neural network, Rule-based pre-processing, Tokenization, Normalization (regex, rules), Binary classification	N/A	materials science articles (abstracts), Hand-annotated abstracts (NER training), JSON files (normalization), Extracted entities, Labeled abstracts (document selection), Dev/Test sets (NER)	Over **3.27 million abstracts**, **800 hand-annotated abstracts** (NER train), 1094 labeled abstracts (doc select), Dev/Test sets size N/A.	(abstracts), NER, Entity normalizati n, Document selection (binary classificati n), Extracting entity type (MAT, SPL, DSC, PRO, APL, SMT, CMT), Tokenizati n, Normalizin chemical formulae
Named entity recognition in chemical patents using ensemble of contextual language models	Contextualized language models, BERT-based architecture, bert-base/large-cased/uncased, ChemBERTa-1 (RoBERTa)	Fine-tuned, Token classification	BERT-base/large, ChemBERTa-1	Corpus of chemical patents (ChEMU lab), Examples of ChEMU NER task, Corpus for ChemBERTa-1 pre-training (SMILES), Corpus for BERT pre-training	Corpus size N/A, Examples size N/A. ChemBERTa-1 pre-training: 100k SMILES. BERT pre-training: large corpus.	NER in chemical patents, Classifying tokens, Extracting informatio Entities: example label, compound types, reaction roles, physical parameter
						Extracting high-qualit chemical reaction

Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature	LLMs, GPT-3.5/4 (OpenAI), GPT-4, Open-sourced LLMs, Llama-2, GPT-3.5-turbo	Prompt-tuning, Manual inspection, Fine-tuning, Verification (fact-based), Prompt engineering, Agent-based, LoRA	7b, 13b (Llama-2), GPT-3.5/4/3.5-turbo/4.0	USPTO database, RXNFP dataset, Curated dataset, Test samples	USPTO large, RXNFP classified 834 classes, Sampled 834 points for prompt-tuning/curation, Curated dataset 8:2 split (size N/A), Test samples size N/A	dataset, Extracting reaction informatio Structurin text, Generating summaries Compiling info, Converting unstructure d to structured, Extracting chemical formulas/reactant names, Extracting reaction data (JSON), Verifying extracted info
	Neural	Manual		SOFC-Exp		IE in materials science, Extracting SOFC experiment info.

The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain	Neural networks, BiLSTM, BERT-based, SciBERT, BERT-base, CRF (baseline), BiLSTM with attention, Combinations of embeddings	Neural network based, Binary sentence classification, Retrieval, Semantic-role-labeling (Slot filling), 5-fold cross validation	BERT-base, SciBERT	SOFC-Exp Corpus, Annotated scholarly articles, Annotation test set, Development set, Synthesis Procedures dataset (comparison)	**45 open-access scientific publications**, Synthesis Procedures: **230 annotated procedures**.	info, Identifying experiment sentences (detection), Identifying materials/ values/ devices (entity mention detection/ typing), Slo filling (experimen slots), Entit extraction (Synthesis Procedures dataset)