Green Chemical Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Green Chemical Engineering

journal homepage: www.keaipublishing.com/en/journals/green-chemical-engineering

Article

Unlocking deep eutectic solvent knowledge through a large language model-driven framework and an interactive AI agent

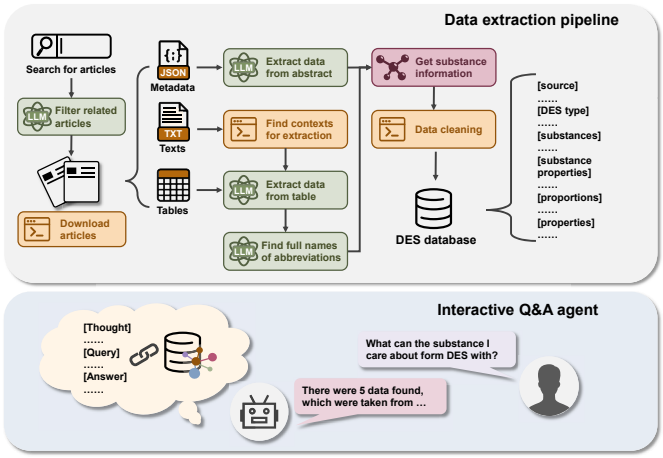
Xiting Peng [a,1], Yi Shen Tew [a,1], Kai Zhao [a,1], Chi Wang [a], Ren'ai Li [c], Shanying Hu [a], Xiaonan Wang [a,b,*]

[a] Department of Chemical Engineering, Tsinghua University, Beijing, 100084, China
[b] Institute for Carbon Neutrality, Tsinghua University, Beijing, 100084, China
[c] Jiangsu Co-innovation Center for Efficient Processing and Utilization of Forest Resources, Jiangsu Provincial Key Lab Pulp & Paper Science and Technology, Nanjing Forestry University, Nanjing, 210037, China

HIGHLIGHTS

- An LLM-driven framework was developed for automated extraction of DES-related data.
- Extraction of 34,027 records and 9,215 unique formulations from 14,602 articles was achieved with over 90% accuracy.
- An AI agent was integrated with a graph-based retrieval system to enable interactive querying.
- A structured DES knowledge base was constructed to accelerate formulation discovery in green chemistry.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Artificial intelligence
Large language model
Deep eutectic solvents
Text mining

ABSTRACT

Artificial intelligence (AI) is playing an important role in advancing green chemical engineering, while the lack of data remains a primary challenge in many fields. Deep eutectic solvents (DESs) are a promising alternative to traditional organic solvents. However, the exploration of new DES formulations has long been constrained by trial-and-error research methods, a preference for familiar formulations, and a lack of easily accessible DES databases. This study proposes a framework driven by large language models (LLMs) for accurately and efficiently extracting data in the DES field, accelerating knowledge discovery. By coordinating LLMs and tools through predefined code paths, we extracted 34,027 data records and 9,215 unique DES formulations from 14,602 research articles, achieving an accuracy of over 90%, thereby creating a comprehensive domain knowledge base. An LLM-driven interactive agent has been deployed on an online platform, further facilitating

* Corresponding author. Department of Chemical Engineering, Tsinghua University, Beijing, 100084, China.
E-mail address: wangxiaonan@tsinghua.edu.cn (X. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.gce.2025.05.006
Received 15 April 2025; Received in revised form 20 May 2025; Accepted 27 May 2025

access to this structured data and enabling researchers to overcome data limitations and accelerate the discovery of new DES formulations.

1. Introduction

The growing demand for sustainability has positioned green chemistry as one of the key directions for the future development of chemical engineering [1]. As foundational components in chemical processes, organic solvents are widely used but often suffer from drawbacks such as high volatility, toxicity, and poor environmental compatibility [2]. In contrast, deep eutectic solvents (DESs) have emerged as a promising class of alternative solvents, offering advantages such as biodegradability, low toxicity, low volatility, and recyclability [3]. Typically composed of two or more components, DESs exhibit highly tunable properties, earning them the designation of "designer solvents" [4]. These features have facilitated their broad applications in various fields, including extraction [5,6], electrochemistry [7], catalysis [8], and materials synthesis [9–11].

The innovation of DES formulations often leads to enhanced solvent performance and expanded application scopes [12,13]. However, the exploration of novel DES formulations remains slow relative to the rapid growth in the number of publications and the vast design space inherent to DESs, which has constrained the realization of their full potential. On one hand, experimental studies on DESs are still guided by traditional trial-and-error paradigms, and the selection of components is often empirically driven. Consequently, research tends to concentrate on a narrow set of commonly used formulations. For instance, the analysis by Ayres et al. [14] revealed that five frequently used hydrogen bond donors (HBDs) accounted for 42.5% of all selections in studies on natural DESs. On the other hand, while machine learning (ML) has been introduced to provide novel insights into DES formulation design [15–20], datasets for such efforts are typically constructed by manually collecting data from a limited number of articles that tested a set of DES properties. Among the 33 DES-related ML studies summarized by López-Flores et al. [21], only five built datasets containing more than 200 distinct DES formulations—a small fraction compared to the huge combinatorial possibilities of multi-component DES formulations. The limited representativeness of such datasets reduces the reliability and generalizability of the resulting ML models.

Furthermore, for researchers in this field, information about specific DES formulations is often scattered across different sections of articles with diverse applications and is not always explicitly summarized in prominent parts of the articles. As a result, it is difficult to gather complete knowledge about DESs using traditional keyword searches. The absence of structured, integrated, and accessible DES data hinders research progress and innovation in this area.

Large language models (LLMs) offer new opportunities for extracting chemical knowledge to address the challenges mentioned above [22]. Traditional literature mining methods often rely on rule-based natural language processing techniques [23–26], which require complex coding and domain-specific rule customization. In contrast, pretrained LLMs come equipped with general knowledge across multiple domains and have demonstrated strong performance in tasks such as text summarization, information extraction, and classification [27]. This makes LLMs a versatile solution for domain knowledge mining, offering advantages such as easier workflow setup, better robustness to non-standard formats, and stronger contextual understanding [28]. LLM-based approaches have already shown promising results in the fields of energy storage [29], catalysis [30], materials design [31–33], and environmental science [34,35]. Beyond information extraction, LLMs can also be used to build multi-agent systems, combining tools and LLMs with different roles to act as research assistants [36,37]. The Retrieval-Augmented Generation (RAG) framework allows LLMs to integrate with external knowledge bases, enriching user queries with relevant information and producing more accurate responses—without the need for model fine-tuning [38,39]. A further development of this concept, Graph Retrieval-Augmented Generation (Graph RAG), incorporates graph-structured data to better capture the relationships between pieces of information [40]. This approach improves the relevance and completeness of the retrieved content while reducing redundant or verbose context [41].

In this work, we present an LLM-driven framework for automated data extraction tailored to the DES domain. By designing a workflow that strategically integrates LLMs with code-based tools, we achieve a balance between accuracy, coverage, and efficiency in both time and cost. Using this system, we extracted 34,027 data records and 9,215 unique DES formulations from 14,602 research articles, resulting in a comprehensive domain-specific knowledge base. Furthermore, we linked this database to an interactive question-answering agent (interactive Q&A agent) powered by LLMs under a Graph RAG framework, enabling researchers to interact with the data through natural language. This allows for flexible and in-depth knowledge retrieval, helping researchers overcome informational obstacles, thereby accelerating discovery and innovation in the DES field.

2. LLM-driven data extraction pipeline and interactive Q&A agent for DES

Fig. 1 illustrates the workflow of the DES data extraction and interactive Q&A agent system integrated with an LLM. This workflow is a system that orchestrates LLMs (represented by the green box) and tools (shown in yellow and purple boxes) through predefined code paths. The data extraction pipeline is primarily divided into three parts: article selection and acquisition, DES information extraction, and database content expansion and cleaning. The extracted database is then used to construct an interactive Q&A agent system, enabling researchers in the field to quickly access targeted DES knowledge data.

The first part of the data extraction pipeline is article selection and acquisition. To collect the most comprehensive data, a broad search term "eutectic" was used on Web of Science, excluding reviews and patents, resulting in 34,939 articles. To filter out irrelevant articles from the DES field (such as those related to metallurgy, ceramics, etc.), the titles and abstracts were provided to the LLM to perform a classification task and identify relevant literature. The filtered articles were then downloaded and converted into formatted JSON files containing metadata, body text, figures, and tables using a literature content extraction tool for subsequent use.

In the DES information extraction stage, the pre-survey indicated that relevant information about DES is primarily found in the abstracts and tables. Therefore, LLMs were mainly used to extract data from these two sections. To minimize the extraction of irrelevant content and efficiently save time and costs, a predefined set of keywords (keyword list available in Note S1 of Supporting Information (SI)) was used to filter tables before extraction. Additionally, a code flow was employed to search for sentences referencing the table and retrieve several surrounding sentences as context, which was then provided to the LLM along with the table content. An example of the prompt used for information extraction is available in Fig. 2 and Note S2 of SI. In contrast to some studies where entries are carefully split into smaller components for extraction [31], we chose to extract all relevant information in a single step. This approach was based on our finding that extracting potentially confusing information together could effectively reduce confusion (for example, LLMs may easily confuse the usage temperature of DES with its melting point, but requesting both of them together significantly alleviates this issue). The prompt is mainly composed of

four parts: task description, a list of information to be extracted (with corresponding notes for each item), format examples, and additional common instructions. This modular design of the prompts allows us to easily add or remove extraction tasks as needed.

Additionally, considering that abbreviations are commonly used in tables to represent components or formulations, an extra step was added for extracting full substance names. The LLM was provided with a list of substance names and the context in which the name first appears. For each name, one of the following four options was returned: the full substance name found in the context; ALREADY FULL NAME, indicating that the provided term is already the full substance name; MOLECULAR FORMULA, suggesting that the provided term is a molecular formula; or UNSPECIFIED, indicating that no full name was found, and the term does not appear to be a full name or molecular formula. To ensure that the extracted substance names are linked to specific molecules and to exclude potential extraction errors, PUG REST, the application programming interface (API) provided by PubChem (https://pubchem.ncbi.nlm.nih.gov/), was used to search for the extracted substance names. These were matched with corresponding entries in the PubChem database to retrieve their molecular formula, molecular structure, and basic property information. This strategy helps achieve more accurate identification of the chemical composition of DES formulations.

Finally, database content expansion and cleaning were performed. A detailed summary of the data cleaning steps is provided in Table S1, including exclusion criteria, percentage of affected records, and representative examples of problematic data. To improve the reliability of the database, any data deemed controversial or ambiguous—such as entries with non-standard formats, structural inconsistencies (*e.g.*, mismatches between the number of components and ratios), components lacking recognized full names or corresponding PubChem records—were removed. While extraction errors do occur, their impact is localized, and mitigation strategies have been built into the data processing pipelines. To ensure data consistency, all substance names were standardized to their PubChem names, while retaining their original names as synonyms. Additionally, to enrich the database content, a series of inferences were made, including inferring substance categories based on molecular structure, inferring the DES type based on the categories of components, and inferring the hydrophilicity or hydrophobicity of mixtures based on the octanol-water partition coefficient (*X*log*P*) and topological polar surface area (TPSA) of the components (detailed inference rules are provided in the Notes S3–S5 of SI). The structured entries generated—including data sources, DES types and formulations, and the properties of components and mixtures—were compiled into an organized DES database.

To facilitate access for researchers, the database is also linked to an interactive Q&A agent. Fig. 3 illustrates the detailed framework of the agent. The proposed agentic Graph RAG workflow aims to seamlessly integrate LLMs and graph databases to enhance interaction with the DES knowledge graph. Initially, the supervisor agent assesses the user's query to decide if a simple, language-model-based response suffices or if a knowledge graph query is necessary. If the query requires specialized DES knowledge, the Cypher query agent constructs and executes a Cypher statement against the graph database. The resulting subgraphs or records are validated, with any errors prompting auto-repair mechanisms to adjust the Cypher query. Once the needed information is extracted, the original query and graph data are fed into the answering agent for a well-contextualized and knowledge-grounded response.

To support multi-turn interactions, the system incorporates a basic memory module that preserves context across consecutive questions. After each response, a concise visualization of the relevant nodes and edges is generated to provide an intuitive representation of how concepts or entities are interrelated within the DES knowledge graph (Fig. S1). Through this agentic approach, the workflow integrates language modeling, automated query composition and execution, and data visualization to deliver thorough and contextually rich answers. The prompts for building the agent are available in Note S6.

In addition, to facilitate easy access, we have deployed the DES agent to an online platform, which can be accessed *via* https://des-agent-48p7tmksubzd2svnreglia.streamlit.app/. A description of the DES database structure, more question examples, and detailed instructions for use can be found on the website's user manual.

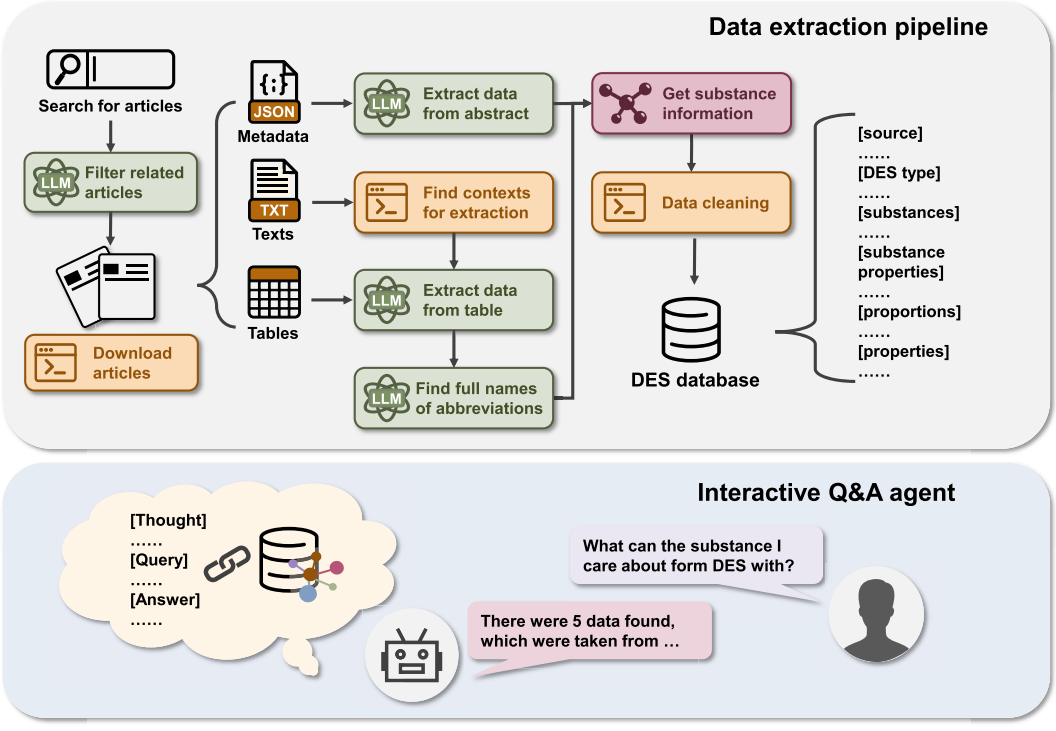


Fig. 1. Overall workflow of the LLM-driven data extraction pipeline and interactive Q&A agent.

3. Evaluation of data extraction result

To assess the reliability of our proposed data extraction pipeline, we tested it on 50 randomly selected articles for paper screening and abstract extraction, and on 60 articles (50 randomly selected and 10 manually selected) with tables for the task of table extraction. Our primary focus was on two aspects: the amount of incorrect data extracted and the amount of data missed. Table 1 summarizes the evaluation results for three major tasks. For the extraction of abstracts and tables, GPT4-turbo achieved accuracies of 89% and 94%, respectively (measured by the number of requests that returned valid data). During the paper screening phase, it successfully excluded 58% of irrelevant articles (15 out of 26), while missing only one relevant article. In comparison, GPT-4o had a lower miss rate but a higher incorrect rate. Our analysis revealed that most errors occurred in paragraphs with inherently ambiguous descriptions. Although GPT-4o demonstrated improved capabilities in handling complex issues, it showed limitations in language comprehension, which affected its ability to identify relevant information. Given the scale of this task, extracting incorrect data is more problematic than missing data. Therefore, we selected GPT4-turbo to perform all extraction tasks.

Based on the data extraction pipeline described in Section 2, 46,548 data points mainly from seven publishers were processed to form the DES database, with 49.2% of the data sourced from papers published by Elsevier (Fig. 4a). After data cleaning, 73.2% of the data was retained as valid unique data, totaling 34,027 records, and redundant data was merged. Additional statistics on the types of DES, the availability of ratios for DES formulations, and the data availability on melting points and effective temperatures are also provided. It is important to note that the classification standards for DES are not yet unified [42–44]. Due to data limitations, we used a relatively coarse classification scheme, dividing mixtures into five types of DES, inorganic salts, and others (details in Note S4 of SI). The results show that type III DESs (15,988 data entries) and type V DESs (10,571 data entries) constitute the majority of the database. The predominance of type V DESs is likely due to their broader applicability and the intensive research focus that they have received in recent years [45]. The availability of ratios for DES formulations is 68.9%. In contrast, the availability of DES melting points is lower (22.1%). By leveraging the ability of LLMs to understand context, we provided inferred melting point ranges based on contextual information (*e.g.*, if the mixture is indicated to be used at room temperature, its melting point is assumed to be below 25 °C; if it is described as solid at a certain temperature, its melting point is assumed to be above that temperature). As a result, 7418 data entries had their melting point ranges effectively supplemented, enriching the database. To provide a clearer understanding of the potential application range of DES, we also included the effective temperature, which refers to the temperature or temperature range at which the mixture is used. 65.1% of the data contains at least one piece of information about its effective temperature. Compared to similar efforts

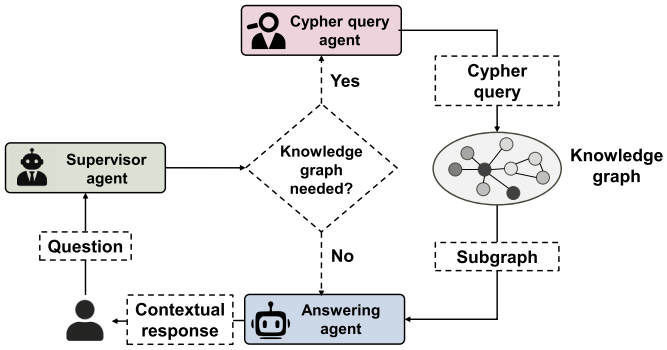


Fig. 3. Detailed framework of interactive Q&A agent.

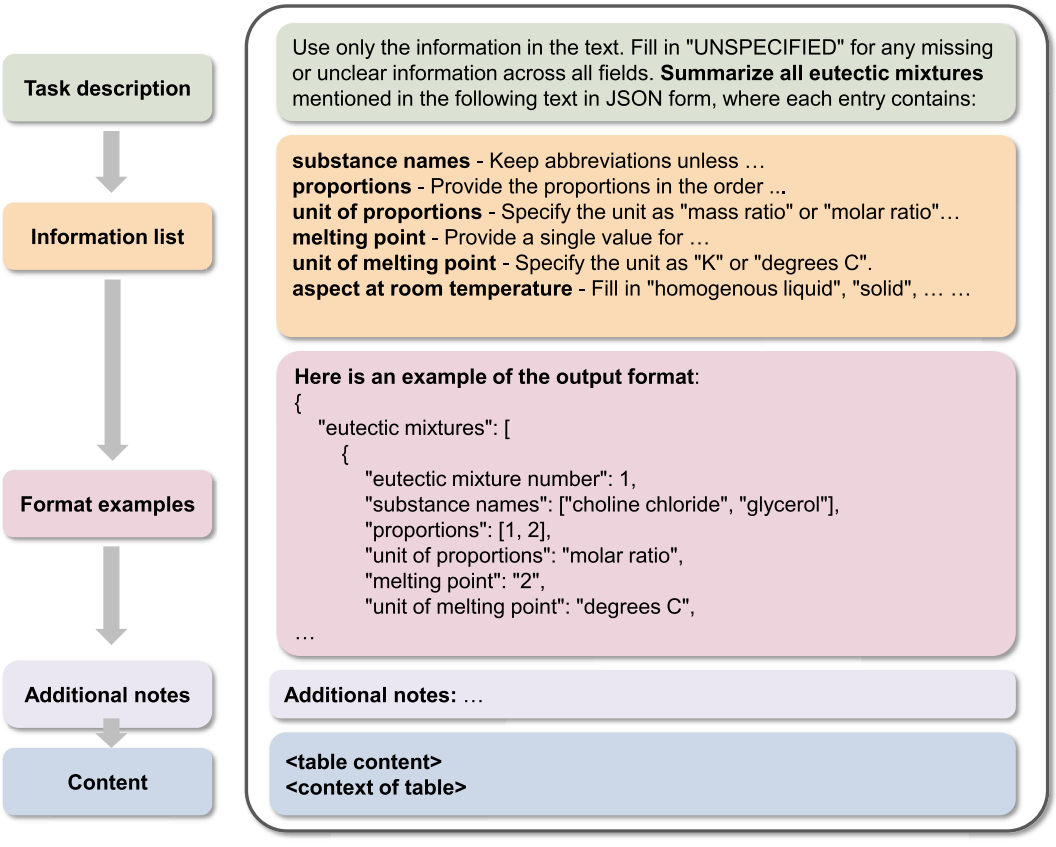


Fig. 2. Schematic illustration of the prompt structure used for DES data extraction.

Table 1
Performance of LLM on three extraction tasks.

Model	Task	Incorrect	Missed	Amount of valid requests
GPT4-turbo	Paper screening	11	1	50
	Abstract extraction	3	2	27
	Table extraction	3	6	50
GPT-4o	Paper screening	14	0	50
	Abstract extraction	5	0	33
	Table extraction	8	5	50

that manually mine DES data [17,20,46–50], the proposed data extraction pipeline demonstrates significant advantages in both the volume and richness of data, covering 9215 unique DES formulations and 7534 entries of single-property data (Fig. 4b). This more comprehensive and accessible DES database will provide stronger support for the exploration of new DES formulations, promoting a paradigm shift in DES research and the discovery of high-performance DES.

To further validate the accuracy of our data extraction results, we conducted a comparative error analysis between our automatically extracted DES database and a manually curated dataset [17]. Specifically, we selected melting point data from the DESignSolvents database—one of the largest available datasets on DES properties [17]. By matching entries based on SMILES and component ratios, we identified 275 overlapping records that contained melting point information in both datasets. We then calculated the absolute differences in reported melting points (Fig. 5). Among these, 195 pairs (70.9%) showed a deviation within ±5 K. The overall mean absolute deviation (MAD) was 9.29 K, while the median absolute deviation was 0.00 K. These results indicate a high level of agreement between automatically and manually extracted data for the majority of entries.

To investigate the sources of larger discrepancies, we manually examined the five cases with the largest melting point deviations, as summarized in Table S2. Interestingly, none of the LLM-extracted values involved any extraction errors. One discrepancy was traced to an incorrect value reported in the original literature, while the remaining four were caused by errors in the manual dataset, in which melting points of pure components or other formulations were mistakenly recorded as those of the DES mixture. These findings highlight the potential for human error in manual curation, especially when processing large datasets, and illustrate the advantage of using automated extraction methods for improved consistency. The high accuracy of LLM-based

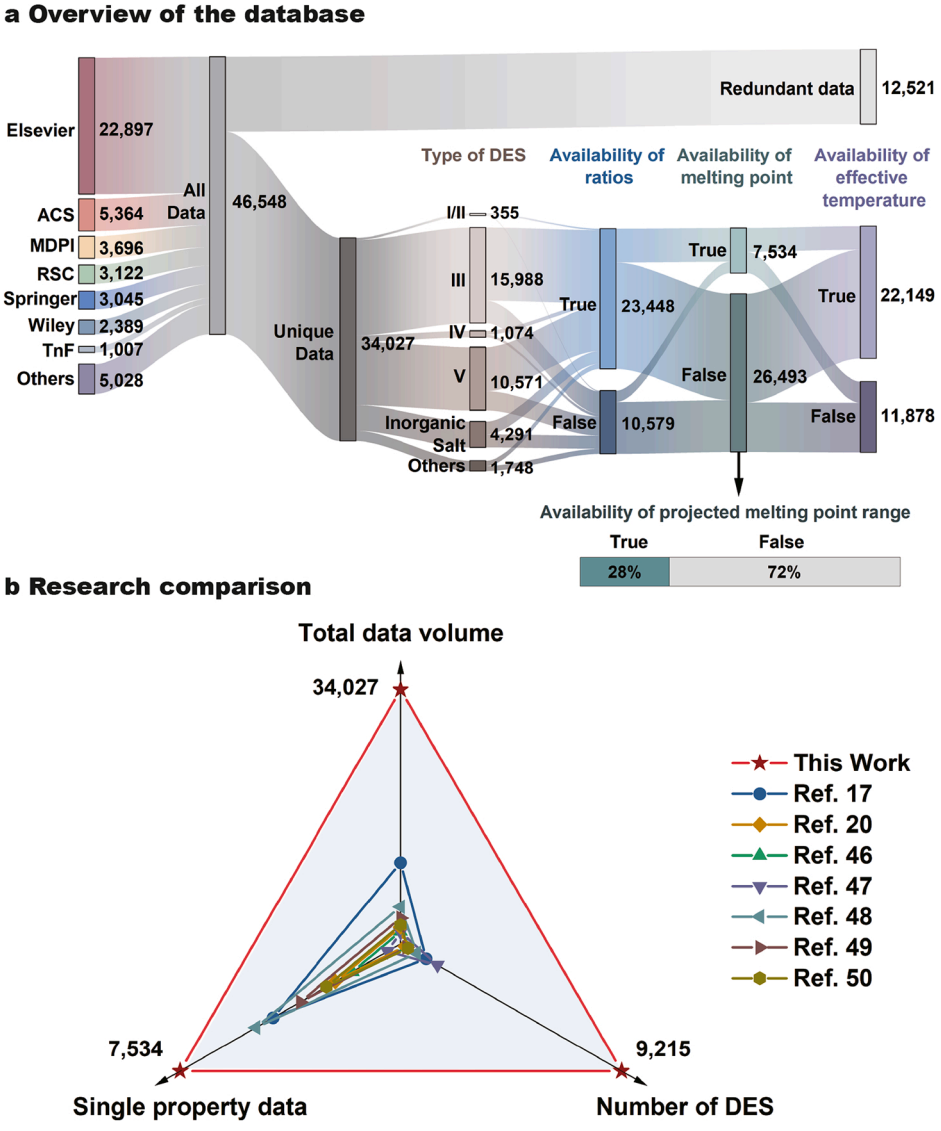


Fig. 4. Results of extracted DES database: (a) overview and (b) comparison of DES database.

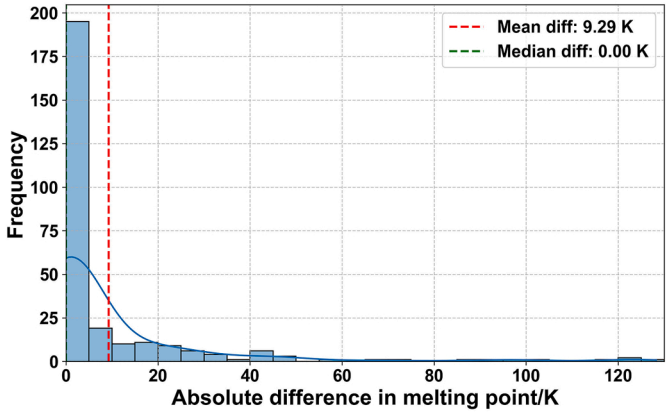


Fig. 5. Distribution of melting point differences between the extracted database in this work compared to DESignSolvents [17].

information extraction is equally evident in energy and materials research cases [51,52].

The presence of incorrect data points can significantly impair downstream applications such as ML model training. This may partly explain why, despite having a larger dataset, the DESignSolvents reported only moderate predictive performance (R^2 = 0.78) [17]. Although further experimental benchmarking is needed to confirm the suitability of our dataset for ML applications, these findings underscore the importance of rigorous data cleaning before model training.

4. Statistical analysis of the extracted DES database

To gain a clear understanding of the extracted data and provide a quantitative insight into the current state of the DES field, we conducted a series of statistical analyses on the database across four dimensions: the components of DES, DES formulations, research frequency, and the properties of mixtures (Fig. 6). The analysis of the components of DES revealed that the database contains 2798 unique chemicals, with 54% being organic molecules, followed by inorganic salts (23%) and organic salts (14%). Additionally, a small number of elements, inorganic non-ionic compounds, and ions are also included (Fig. 6a). Among these components, half do not possess HBD sites, while most components have sites that can act as hydrogen bond acceptors (HBAs) (Fig. 6b–c). This is because, in some molecules, the presence of HBDs is determined by specific functional groups (*e.g.*, R–NH_2, R–OH), but not all molecules contain such functional groups. On the other hand, functional groups containing oxygen (*e.g.*, R-*O*-R′, R–COOH), nitrogen (*e.g.*, R–$CONH_2$), or halogens (*e.g.*, F) can act as HBAs and are commonly found in a wide variety of compounds. Common groups that act as HBDs, such as the carboxyl group in acids, also typically have HBA sites. Most components contain only a small number of HBAs and HBDs, confirming the current trend in the field of using relatively simple small molecules. However, a small fraction of components exhibit higher HBA or HBD capabilities.

TPSA, as an indicator of molecular polarity, exhibits a distribution pattern similar to that of HBA and HBD. Most DES components tend to have low TPSA values or lack polar functional groups, although there are also a small number of components with high TPSA values (Fig. 6d). *X*Log*P* is used to describe the hydrophilicity or hydrophobicity of a substance. When *X*Log*P* < 0, the substance tends to be hydrophilic, while *X*Log*P* > 0 indicates a greater tendency towards hydrophobicity. Most DES components exhibit moderate hydrophilicity or hydrophobicity, but a significant number of components are hydrophobic (Fig. 6e).

Among the 9,215 different DES formulations in the database, the formulations of type V and type III DES, primarily composed of organic molecules and organic salts, are the most diverse (Fig. 6f), reflecting the flexibility of DES formulations. Analysis of the inferred hydrophilicity or hydrophobicity for different formulations (inference rules provided in the Note S5 of SI) shows that type III and type IV DES tend to be more hydrophilic, while type V DES can be adjusted between hydrophilic, hydrophobic, and amphiphilic properties, highlighting the tunability of DES (Fig. 6g). Significant data gaps are observed in the I/II and inorganic salt categories, as the *X*Log*P* value, which indicates the distribution tendency between organic and inorganic phases, is not meaningful for most inorganic salts.

Research frequency analysis reveals that among all DES formulations, the choline chloride and urea-based formulation has attracted the most attention, with 1,944 articles in the database investigating it (Fig. 6h). Of the 14,602 articles collected, 5,719 (39.17%) mention choline chloride, and four of the ten most used substances are chlorides (Fig. 6i). This trend highlights the preference of researchers for a few specific DES formulations. The preference for chlorides is due to chlorine being a strong HBA, making it an excellent candidate for forming DES. Additionally, chloride chemistry is part of a well-established industrial system, with its derivative chemicals being relatively inexpensive and easily accessible.

For the properties of mixtures, Fig. 6j shows the distribution of the average TPSA and *X*Log*P* of the mixtures calculated based on the corresponding values of the constituent substances (Note S5 of SI). Substances with high TPSA values and low *X*Log*P* tend to be more hydrophilic. As seen, although there is a greater variety of substances leaning towards hydrophobicity (Fig. 6e), the number of hydrophilic formulations (*X*Log*P* < 0) is still higher at the mixture level. Fig. 6k presents the distribution of melting points and effective temperatures for the mixtures. Most mixtures show a normal-like distribution centered around room temperature for their melting points, though a significant number of substances have higher melting points (> 200 °C). This highlights a potential limitation in the database: despite several rounds of cleaning, some data may still pertain to formulations that are not of primary interest in the DES field. However, to maintain data comprehensiveness, this subset of data has been retained. For effective temperature, two distinct peaks are observed at 20–30 °C and 80–90 °C, which corroborate the typical usage and preparation temperatures for common DES formulations.

To explore how commonly used DES formulations vary across different application contexts, we conducted a comparative analysis focusing on two emerging application domains: DESs used as electrolytes in batteries and supercapacitors, and DESs applied in valuable metal recovery for battery recycling. These two fields were selected because they both relate to the broader theme of electrochemical energy technologies, which are of increasing interest in the context of sustainable energy and resource management. However, they emphasize different performance requirements: electrolyte applications focus on ionic conductivity and electrochemical stability, while metal recovery prioritizes selective dissolution and complexation efficiency.

Using an LLM-based classification approach, we identified 335 articles related to DES electrolytes and 187 articles related to metal recovery, based on keyword-guided filtering of titles and abstracts. From the articles mapped to our database, we selected the ten most frequently studied DES formulations in each field and analyzed their evolution over time. The cumulative number of studies for each formulation was plotted to illustrate changes in research focus (Fig. 7).

In the electrolyte domain, research began earlier—around 2010—and initially centered on the well-established choline chloride/urea system. This was followed by increased interest in the formulation of choline chloride/ethylene glycol, and later, DESs tailored for specific battery types emerged: *e.g.*, lithium bis((trifluoromethyl)sulfonyl)azanide (LiTFSI)-based systems for lithium-ion batteries, $AlCl_3$-based systems for aluminum batteries, and $Zn(NO_3)_2{\cdot}6H_2O$-based systems for zinc batteries.

By contrast, DESs for valuable metal recovery began to appear more recently, around 2015, driven by growing attention to resource circularity. This field also shows strong reliance on formulations of choline

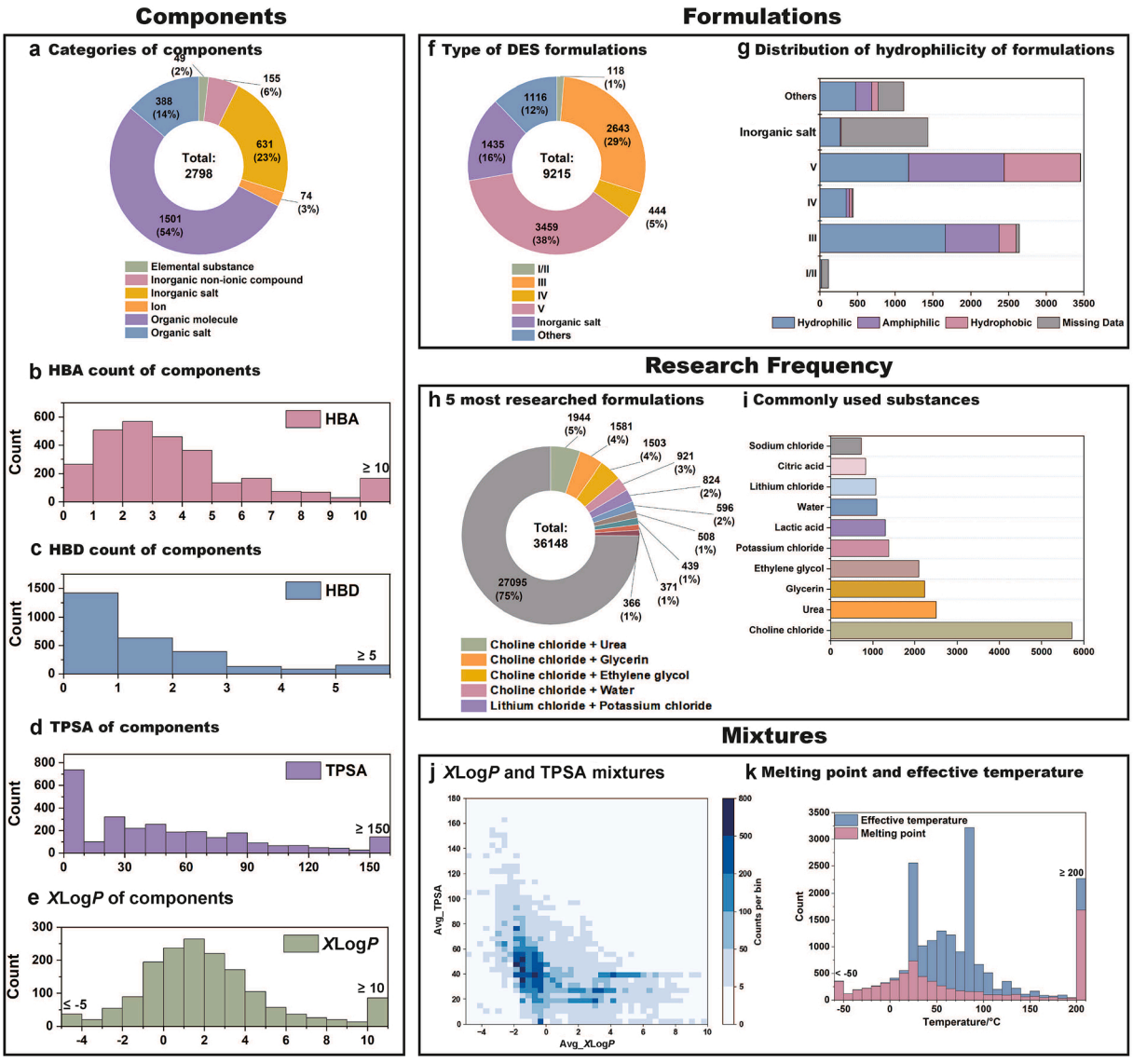


Fig. 6. Statistical analysis of the extracted DES database: (a) categories, (b) HBA count, (c) HBD count, (d) TPSA, and (e) *XLogP* of components; (f) type and (g) distribution of hydrophilicity of DES formulations; the research frequency of 10 most researched (h) DES formulations and (i) substances; (j) average *XLogP* and TPSA of mixtures; (k) melting point and effective temperature of mixtures.

chloride/urea and choline chloride/ethylene glycol, while other frequently used systems often involve choline chloride paired with various organic acids to dissolve metal oxides.

5. Navigating DES knowledge: the role of AI agent in research

To better demonstrate the practical value of the proposed LLM framework, Fig. 8 illustrates a representative query interaction using the DES agent. This example simulates a typical research task, where a user seeks to identify DES formulations that contain a substance of interest and meet specific property requirements. The user then examines their components and traces the original literature sources. Such a multi-step query chain, if conducted manually, would require extensive time navigating across multiple literature sources and may also miss some of the relevant studies. In contrast, the agent accomplishes this task through structured natural language queries mapped to Cypher statements over a curated knowledge graph (Fig. 3), offering substantial improvements in efficiency, precision, and accessibility, especially for exploring less commonly studied or complex formulations.

Each query cycle shown in Fig. 8 involves contextual interpretation of the user's intent, followed by graph-based retrieval of verified information. A more detailed illustration of this example can be found in Note S7 of SI, and a demonstration video is also attached within the SI. Unlike generic LLM-based responses which may suffer from hallucination or unverifiable content, the DES agent operates strictly over a structured and traceable database, ensuring that all outputs are source-linked data. This hybrid approach combines the flexibility of LLMs with the reliability of database-supported reasoning, offering both speed and scientific credibility for DES discovery workflows, and laying the foundation for automated laboratories in the DES field [53].

While the current agent does not perform predictive or statistical tasks such as property estimation or candidate mixture evaluation, it serves as a powerful tool for pre-screening and hypothesis generation. Rapidly retrieving formulations that meet specific criteria and providing their complete structural and bibliographic context helps researchers efficiently narrow down viable experimental targets. Future work may build upon this foundation by expanding the database to include a broader range of physicochemical and functional properties—such as viscosity, ionic conductivity, and thermal stability—and by embedding lightweight data analysis tools, including property distribution

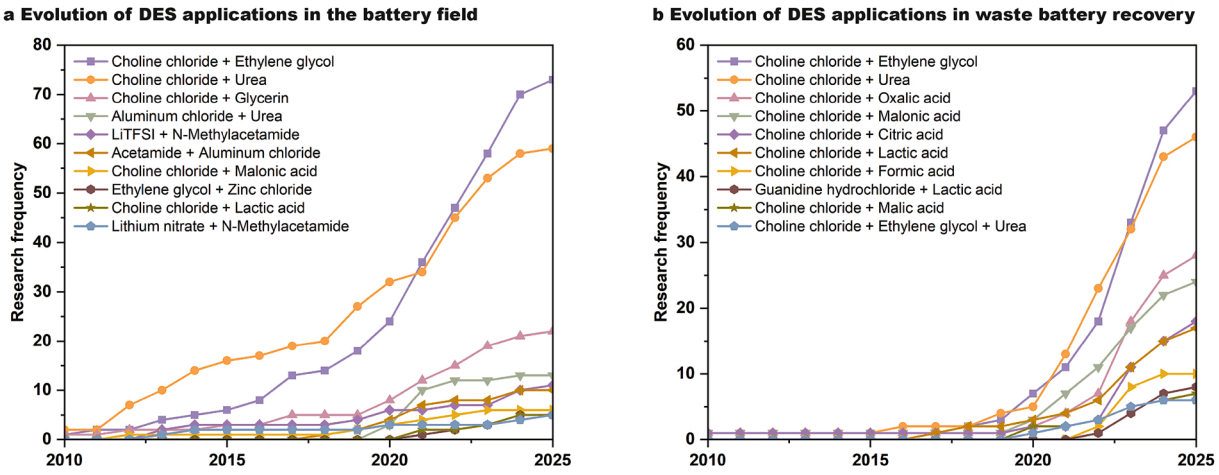


Fig. 7. Temporal evolution of the ten most frequently studied DES formulations in two emerging application domains: (a) as electrolytes in batteries and supercapacitors and (b) for valuable metal recovery in battery recycling. The y-axis represents the cumulative number of studies involving each formulation up to a given year.

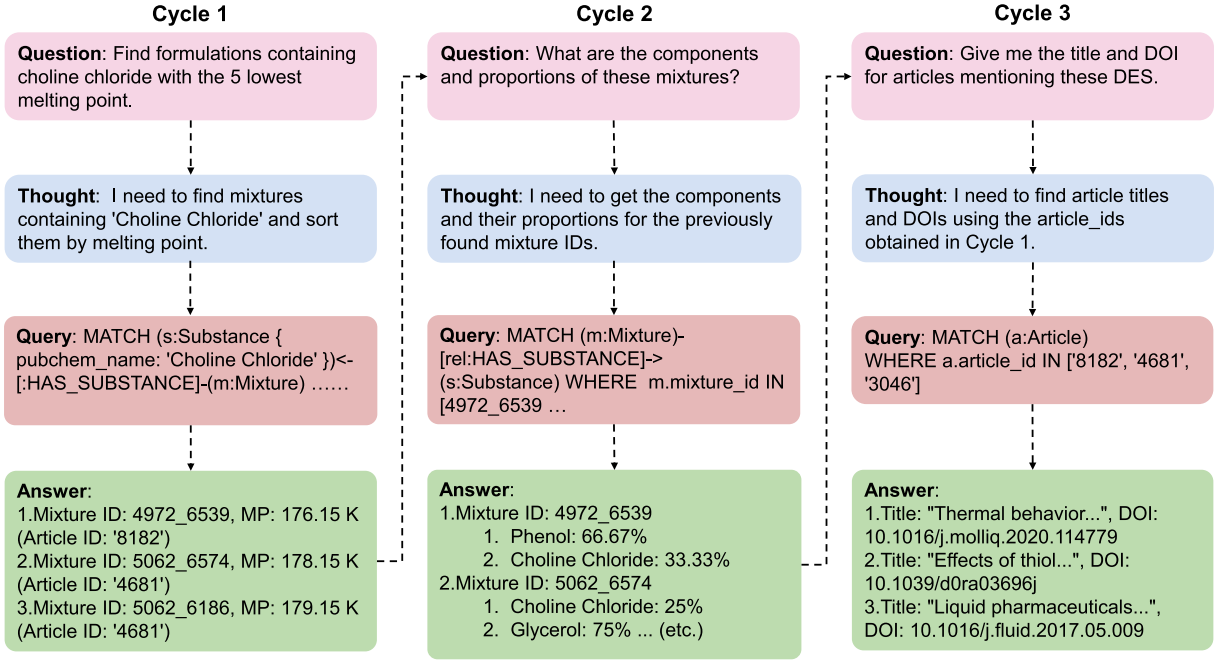


Fig. 8. Example three-turn conversation with the DES agent.

visualization, trend analysis, or unsupervised clustering, to support preliminary interpretation and insight generation. Moreover, integrating predictive ML models could further enhance the framework's utility, evolving the agent into a proactive research assistant that supports a wider range of scientific workflows—from data collection and hypothesis formulation to experimental planning and iterative discovery.

6. Conclusion

In this work, we presented a framework for extracting and retrieving knowledge on DESs using LLMs. The framework combines automated data extraction, content curation, and interactive querying, significantly enhancing the accessibility of DES knowledge. By leveraging this framework, we created a comprehensive DES database containing 34,027 data records from 14,602 research articles, incorporating diverse formulations and properties. The integration of a Graph RAG-based AI agent facilitates more efficient data retrieval, helping researchers quickly access relevant information and overcome challenges related to fragmented and scattered knowledge. This work contributes to a paradigm shift in DES research by providing a more systematic, automated, and accessible way to explore DES properties, formulations, and their potential applications. Moving forward, this approach can be extended to other areas of green chemistry, aiding the discovery and development of sustainable materials and solvents.

CRediT authorship contribution statement

Xiting Peng: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Yi Shen Tew:** Writing – original draft, Software, Methodology, Investigation, Data curation. **Kai Zhao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Chi Wang:** Software. **Ren'ai Li:** Writing –

review & editing, Investigation. **Shanying Hu:** Supervision. **Xiaonan Wang:** Writing – review & editing, Resources, Project administration, Funding acquisition.

Data availability

The extracted database can be accessed through the DES agent constructed in this work: https://des-agent-48p7tmksubzd2svnreglia.streamlit.app/. All code for DES agent in this study is available on GitHub (https://github.com/hyperdrive00/DES-Agent).

Declaration of competing interests

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

Acknowledgments

This work is supported by the Carbon Neutrality and Energy System Transformation (CNEST) Program led by Tsinghua University, Tsinghua University Initiative Scientific Research Program, and the Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCXNLZCXM-E7).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gce.2025.05.006.

References

[1] M. Poliakoff, P. Licence, Green chemistry, Nature 450 (2007) 810–812.
[2] F.M. Perna, P. Vitale, V. Capriati, Deep eutectic solvents and their applications as green solvents, Curr. Opin. Green Sustainable Chem. 21 (2020) 27–33.
[3] C. Florindo, L.C. Branco, I.M. Marrucho, Quest for green-solvent design: from hydrophilic to hydrophobic (deep) eutectic solvents, ChemSusChem 12 (2019) 1549–1559.
[4] J.D. Mota-Morales, R.J. Sánchez-Leija, A. Carranza, J.A. Pojman, F. del Monte, G. Luna-Bárcenas, Free-radical polymerizations of and in deep eutectic solvents: green synthesis of functional materials, Prog. Polym. Sci. 78 (2018) 139–153.
[5] Y. Dai, G.-J. Witkamp, R. Verpoorte, Y.H. Choi, Tailoring properties of natural deep eutectic solvents with water to facilitate their applications, Food Chem. 187 (2015) 14–19.
[6] Q. Xia, C. Chen, Y. Yao, J. Li, S. He, Y. Zhou, T. Li, X. Pan, Y. Yao, L. Hu, A strong, biodegradable and recyclable lignocellulosic bioplastic, Nat. Sustain. 4 (2021) 627–635.
[7] A.P. Abbott, Deep eutectic solvents and their application in electrochemistry, Curr. Opin. Green Sustainable Chem. 36 (2022) 100649.
[8] M. Pätzold, S. Siebenhaller, S. Kara, A. Liese, C. Syldatk, D. Holtmann, Deep eutectic solvents as efficient solvents in biocatalysis, Trends Biotechnol. 37 (2019) 943–959.
[9] D. Yu, D. Jiang, Z. Xue, T. Mu, Deep eutectic solvents as green solvents for materials preparation, Green Chem. 26 (2024) 7478–7507.
[10] X. Li, J. Liu, Q. Guo, X. Zhang, M. Tian, Polymerizable deep eutectic solvent-based skin-like elastomers with dynamic schemochrome and self-healing ability, Small 18 (2022) 2201012.
[11] K. Zhao, K. Zhang, R. Li, P. Sang, H. Hu, M. He, A very mechanically strong and stretchable liquid-free double-network ionic conductor, J. Mater. Chem. A 9 (2021) 23714–23721.
[12] P. Ding, K. Zhao, R. Li, P. Sang, X. Liang, K. Zhang, X. Liu, G. Chen, M. He, Highly ionic conductive and transparent self-healing lithium salt elastomers based on eutectic strategy, Chem. Mater. 34 (2022) 10320–10328.
[13] R. Li, C. Su, M. Li, Y. Cao, Innovative green synthesis of hydrophobic covalent networks using ethyl cellulose/thymol eutectic systems, Green Chem. 26 (2024) 10529–10537.
[14] L.B. Ayres, G. Weavil, M. Alhoubani, B.G.S. Guinati, C.D. Garcia, Big data for a deep problem: understanding the formation of NADES through comprehensive chemical analysis and RDKit, J. Mol. Liq. 389 (2023) 122891.
[15] M. Mohan, O. Demerdash, B.A. Simmons, J.C. Smith, M.K. Kidder, S. Singh, Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network, Green Chem. 25 (2023) 3475–3492.
[16] M. Mohan, O.N. Demerdash, B.A. Simmons, S. Singh, M.K. Kidder, J.C. Smith, Physics-based machine learning models predict carbon dioxide solubility in chemically reactive deep eutectic solvents, ACS Omega 9 (2024) 19548–19559.
[17] V. Odegova, A. Lavrinenko, T. Rakhmanov, G. Sysuev, A. Dmitrenko, V. Vinogradov, DESignSolvents: an open platform for the search and prediction of the physicochemical properties of deep eutectic solvents, Green Chem. 26 (2024) 3958–3967.
[18] A.T.N. Fajar, T. Hanada, A.D. Hartono, M. Goto, Estimating the phase diagrams of deep eutectic solvents within an extensive chemical space, Commun. Chem. 7 (2024) 27.
[19] T. Lemaoui, A. Boublia, A.S. Darwish, M. Alam, S. Park, B.-H. Jeon, F. Banat, Y. Benguerba, I.M. AlNashef, Predicting the surface tension of deep eutectic solvents using artificial neural networks, ACS Omega 7 (2022) 32194–32207.
[20] L.-Y. Yu, G.-P. Ren, X.-J. Hou, K.-J. Wu, Y. He, Transition state theory-inspired neural network for estimating the viscosity of deep eutectic solvents, ACS Cent. Sci. 8 (2022) 983–995.
[21] F.J. López-Flores, C. Ramírez-Márquez, J.B. González-Campos, J.M. Ponce-Ortega, Machine learning for predicting and optimizing physicochemical properties of deep eutectic solvents: review and perspectives, Ind. Eng. Chem. Res. 64 (2025) 3103–3117.
[22] M.C. Ramos, C.J. Collison, A.D. White, A review of large language models and autonomous agents in chemistry, Chem. Sci. 16 (2025) 2514–2572.
[23] M.C. Swain, J.M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, J. Chem. Inf. Model. 56 (2016) 1894–1904.
[24] J. Mavračić, C.J. Court, T. Isazawa, S.R. Elliott, J.M. Cole, ChemDataExtractor 2.0: autopopulated ontologies for materials science, J. Chem. Inf. Model. 61 (2021) 4280–4289.
[25] Z. Jensen, E. Kim, S. Kwon, T.Z.H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, A machine learning approach to zeolite synthesis enabled by automatic literature data extraction, ACS Cent. Sci. 5 (2019) 892–899.
[26] S. Park, B. Kim, S. Choi, P.G. Boyd, B. Smit, J. Kim, Text mining metal–organic framework papers, J. Chem. Inf. Model. 58 (2018) 244–251.
[27] B. Min, H. Ross, E. Sulem, A.P. Ben Veyseh, T.H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing *via* large pre-trained language models: a survey, ACM Comput. Surv. 56 (2024) 1–40.
[28] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M.V. Gil, S. Miret, C.T. Koch, J. A. Márquez, K.M. Jablonka, From text to insight: large language models for chemical data extraction, Chem. Soc. Rev. 54 (2025) 1125–1150.
[29] Y. Na, J.J. Kim, C. Park, J. Hwang, C. Kim, H. Lee, J. Lee, Advanced scientific information mining using LLM-driven approaches in layered cathode materials for sodium-ion batteries, Mater. Adv. 6 (2025) 2543–2548.
[30] H. Liu, H. Yin, Z. Luo, X. Wang, Integrating chemistry knowledge in large language models *via* prompt engineering, Synth. Syst. Biotechnol. 10 (2025) 23–38.
[31] M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nat. Commun. 15 (2024) 1569.
[32] Y. Kang, W. Lee, T. Bae, S. Han, H. Jang, J. Kim, Harnessing large language models to collect and analyze metal–organic framework property data set, J. Am. Chem. Soc. 147 (2025) 3943–3958.
[33] Z. Zheng, O. Zhang, C. Borgs, J.T. Chayes, O.M. Yaghi, ChatGPT chemistry sssistant for text mining and the prediction of MOF synthesis, J. Am. Chem. Soc. 145 (2023) 18048–18062.
[34] F. Cheng, J. Huang, H. Li, B.I. Escher, Y. Tong, M. König, D. Wang, F. Wu, Z. Yu, B. W. Brooks, J. You, Text mining-based suspect screening for aquatic risk assessment in the big data era: event-driven taxonomy links chemical exposures and hazards, Environ. Sci. Technol. Lett. 10 (2023) 1004–1010.
[35] W. Liang, W. Su, L. Zhong, Z. Yang, T. Li, Y. Liang, T. Ruan, G. Jiang, Comprehensive characterization of oxidative stress-modulating chemicals using GPT-based text mining, Environ. Sci. Technol. 58 (2024) 20540–20552.
[36] D.A. Boiko, R. Macknight, B. Kline, G. Gomes, Autonomous chemical research with large language models, Nature 624 (2023) 570–578.
[37] Y. Kang, J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, Nat. Commun. 15 (2024) 4705.
[38] J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S.G. Rodriques, A.D. White, Paperqa: retrieval-augmented generative agent for scientific research, arXiv preprint, arXiv: 2312.07559 (2023).
[39] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on RAG meeting LLMs: towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6491–6501.
[40] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: a survey, arXiv preprint, arXiv:2408.08921 (2024).
[41] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R.A. Rossi, S. Mukherjee, X. Tang, Retrieval-augmented generation with graphs (graphrag), arXiv preprint, arXiv:2501.00309 (2024).
[42] Y. Bao, Y. Wang, C. Yan, Z. Xue, Deep eutectic solvents for fractionation and valorization of lignocellulose, Green Chem. Eng. 6 (2025) 21–35.
[43] Y. Chen, Z. Yu, Low-melting mixture solvents: extension of deep eutectic solvents and ionic liquids for broadening green solvents and green chemistry, Green Chem. Eng. 5 (2024) 409–417.
[44] A. Khalid, S. Tahir, A.R. Khalid, M.A. Hanif, Q. Abbas, M. Zahid, Breaking new grounds: metal salts based-deep eutectic solvents and their applications—a comprehensive review, Green Chem. 26 (2024) 2421–2453.
[45] D.O. Abranches, M.A.R. Martins, L.P. Silva, N. Schaeffer, S.P. Pinho, J.A. P. Coutinho, Phenolic hydrogen bond donors in the formation of non-ionic deep eutectic solvents: the quest for type V DES, Chem. Commun. 55 (2019) 10253–10256.

[46] A.K. Lavrinenko, I.Y. Chernyshov, E.A. Pidko, Machine learning approach for the prediction of eutectic temperatures for metal-free deep eutectic solvents, ACS Sustainable Chem. Eng. 11 (2023) 15492–15502.
[47] K.A. Omar, R. Sadeghi, Database of deep eutectic solvents and their physical properties: a review, J. Mol. Liq. 384 (2023) 121899.
[48] M. Mohan, K.D. Jetti, M.D. Smith, O.N. Demerdash, M.K. Kidder, J.C. Smith, Accurate machine learning for predicting the viscosities of deep eutectic solvents, J. Chem. Theor. Comput. 20 (2024) 3911–3926.
[49] D.M. Makarov, A.M. Kolker, Viscosity of deep eutectic solvents: predictive modeling with experimental validation, Fluid Phase Equilib. 587 (2025) 114217.
[50] A. Roosta, R. Haghbakhsh, A. Rita C. Duarte, S. Raeissi, Deep eutectic solvent viscosity prediction by hybrid machine learning and group contribution, J. Mol. Liq. 388 (2023) 122747.
[51] G. Buster, P. Pinchuk, J. Barrons, R. McKeever, A. Levine, A. Lopez, Supporting energy policy research with large language models: a case study in wind energy siting ordinances, Energy AI 18 (2024) 100431.
[52] S. Li, Z. Huang, Y. Li, S. Deng, X.E. Cao, Methodology for predicting material performance by context-based modeling: a case study on solid amine CO_2 adsorbents, Energy AI 20 (2025) 100477.
[53] J. Wang, G. Li, Full-scale modeling of chemical experiments, Smart Mol. 2 (2024) e20230010.