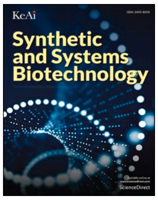
Synthetic and Systems Biotechnology 10 (2025) 23–38

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology



Original Research Article

Integrating chemistry knowledge in large language models via prompt engineering

Hongxuan Liu [a,1], Haoyu Yin [a,1], Zhiyao Luo [c], Xiaonan Wang [a,b,*]

[a] Department of Chemical Engineering, Tsinghua University, Beijing, 100084, China
[b] Key Laboratory for Industrial Biocatalysis, Ministry of Education, Tsinghua University, Beijing, 100084, China
[c] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Old Road Campus Research Building, Headington, Oxford, OX3 7DQ, United Kingdom

ABSTRACT

This paper presents a study on the integration of domain-specific knowledge in prompt engineering to enhance the performance of large language models (LLMs) in scientific domains. The proposed domain-knowledge embedded prompt engineering method outperforms traditional prompt engineering strategies on various metrics, including capability, accuracy, F1 score, and hallucination drop. The effectiveness of the method is demonstrated through case studies on complex materials including the MacMillan catalyst, paclitaxel, and lithium cobalt oxide. The results suggest that domain-knowledge prompts can guide LLMs to generate more accurate and relevant responses, highlighting the potential of LLMs as powerful tools for scientific discovery and innovation when equipped with domain-specific prompts. The study also discusses limitations and future directions for domain-specific prompt engineering development.

1. Introduction

The rapid advancement in artificial intelligence (AI) has significantly propelled its integration into natural science, specifically chemistry and material science. Virtual screening contains thresholds determination and properties labeling, which can exhaust known design space [1] and guide experimental explorations [2]. Designing thresholds requires thorough domain insights but the rate-determining step in virtual screening is labeling the data. Early applications of AI in science were focused on properties predictions (e.g. formation energy [3], selectivity & permeability of membranes [4–6], protein structures [7,8] and batteries' life [9,10]). As machine learning advances, more variants of artificial neural networks enabled AI to handle information in complex modal and solve more sophisticated problems in computational chemistry. For example, MLP (multilayer perceptron) based machine learning potentials for molecular dynamics [11,12], GNN based DFT (density functional theory) functionals [13,14], CNN based electron microscope images processing [15,16]. However, traditional high-throughput virtual screening is limited to known molecules or materials. The emergence of AI in inverse design emphasizes the need for innovative models that can assist experts in discovering new structures [17]. Models containing generating and predicting enable de novo design of molecules [18,19], drugs [20] and proteins [21,22].

A key challenge in applying AI to science is the lack of experimental data, which is often costly and time-consuming to gather. For example, one manual hydrogen evolution measurement experiment takes half a day and an average time of proceeding requires about several months [23]. Even with automated instruments, perovskite crystal formation experiments take 20 months to obtain 8470 datapoints [24].

Overcoming the 'small data' challenge is basic but essential. Among tremendous approaches towards improving learning efficiency, large language models (LLMs) open a new channel for more efficient virtual screening apart from conventional methods, such as high-throughput computational methods [25–27], autonomous wet experiments [28, 29], and data efficient algorithms (e.g. Bayesian optimization [30,31] and active learning [32]). LLMs are capable of processing and analyzing vast data amounts, which have notably advanced in addressing challenges like zero-shot reasoning, enabling them to handle tasks they have not been explicitly trained for. They also excel in incorporating domain knowledge across various fields and providing explanations in natural language, thereby enhancing their adaptability and accessibility. The LLM based AI agents [33] and pre-trained foundation models [34,35] are considered as the next generation of AI scientific assistants.

Prompt quality affects LLMs' outputs significantly, many studies

Peer review under responsibility of KeAi Communications Co., Ltd.

[*] 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.synbio.2024.07.004
Received 8 April 2024; Received in revised form 8 July 2024; Accepted 20 July 2024
Available online 24 July 2024

focus on well-defined prompts for general purposes (e.g., chain of thoughts reasoning [36], few shots learning [37]), known as prompt engineering [38–40]. Enhancing LLMs for specific fields typically involves fine-tuning, which can be complex and costly for those outside AI community [41]. Although there are already some domain-specific LLMs, they have not yet achieved the stability of general-purpose models like ChatGPT, leading many to focus on how to effectively utilize ChatGPT. Considering LLMs' remarkable learning abilities, strategic prompting or directing the LLM with specific instructions could be an effective alternative. However, current prompt engineering is mostly focused on general conditions such as academic writing [42] and science popularizing [43]. For experts in non-AI disciplines, the true value of these models lies in their domain-specific expertise, rather than their general capabilities. The absence of prompt engineering for specific areas makes LLMs user-unfriendly, especially for experimental chemists and material scientists.

Our paper studies the overlooked gap in AI for chemistry and materials science, including small molecules, crystal materials and protein enzymes, highlight the importance of prompting to researchers off-the-shelve LLM. Our investigation shows the critical need for solutions that combine AI's generative capabilities with detailed materials science insights, aiming to enhance model applicability and to address domain-specific challenges across various research areas.

In this article, we introduced "domain-knowledge embedded prompt engineering" as a novel approach to enhance LLM performance in specialized areas, as depicted in Fig. 1. First, we created a set of domain-specific datasets for the first time, supplementing the existing public datasets. Second, we developed and tested specific prompts for various tasks in three examples extracted from chemistry, materials science, and biology. Third, we combined the general methods of the computer science community for comparison, validating that the approach is correct. This approach aligns with desired outcomes and involves developing appropriate evaluation metrics. We also addressed the issue of LLMs generating inaccurate or 'hallucinated' responses and designed strategies to mitigate this. Last, through a case study, we demonstrated how our prompting strategies can address specific challenges in these fields. Overall, we showed that domain-knowledge embedded prompt engineering offers a cost-effective and efficient way to leverage the potential of LLMs.

2. Methods

In this chapter, we first introduce the construction of tasks from three domains: organic small molecules, enzymes and crystal materials, and the answer evaluation scheme for numerical and verbal tasks (See Section 2.1). We then formulate these tasks of domain question answering to an LLM question answering problem (See Section 2.2) and introduce various existing prompt engineering methods to address these tasks (See Section 2.3). Finally, we put forward our domain-knowledge embedded prompt engineering method (See Section 2.3).

2.1. Dataset construction and answer evaluation scheme

In task construction process, each of the three material categories (small molecule, enzyme and crystal material), holds significant relevance in academic research and practical applications. Organic small molecules are commonly utilized in pharmacy [44], while enzymes play a critical role in biocatalysis [45,46], and crystalline materials are essential in semiconductor technology and photovoltaic devices [47,48]. While mainstream benchmark datasets such as MMLU [49], Big-Bench [50] and GSM8k [51] have been widely applied to LLM performance evaluation, the composition of these datasets are usually generic math or reasoning questions, lacking a concentrated focus on some specific knowledge domains or subjects. Compared to these datasets, our datasets could provide a more comprehensive evaluation of LLM's performance (using different prompt engineering methods) on specific chemistry domains.

We collected and curated a dataset of 1280 questions and

Table 1
The composition of prompt engineering datasets.

Datasets	Tasks	Number of Molecules	Number of Tasks
Crystal Material	Space Group Number, Lattice Angle (α,β,γ), Lattice Vector (a,b, c), Density, Formation Energy, Energy Above Hull, Stability, Band Gap, Direct Gap, Metallic, Total Magnetization, Ordering	40	**640**
Organic Small Molecule	Molecular Formula, Melting Point, Density, Solubility, Molecular Weight, H-bond Acceptors, H-bond Donors, LogP, Drugability	40	**360**
Enzyme	Category, Substrate, Product, Active Site, Biological Process, Number of Amino Acids, Ligand	40	**280**
Total	32	120	**1280**

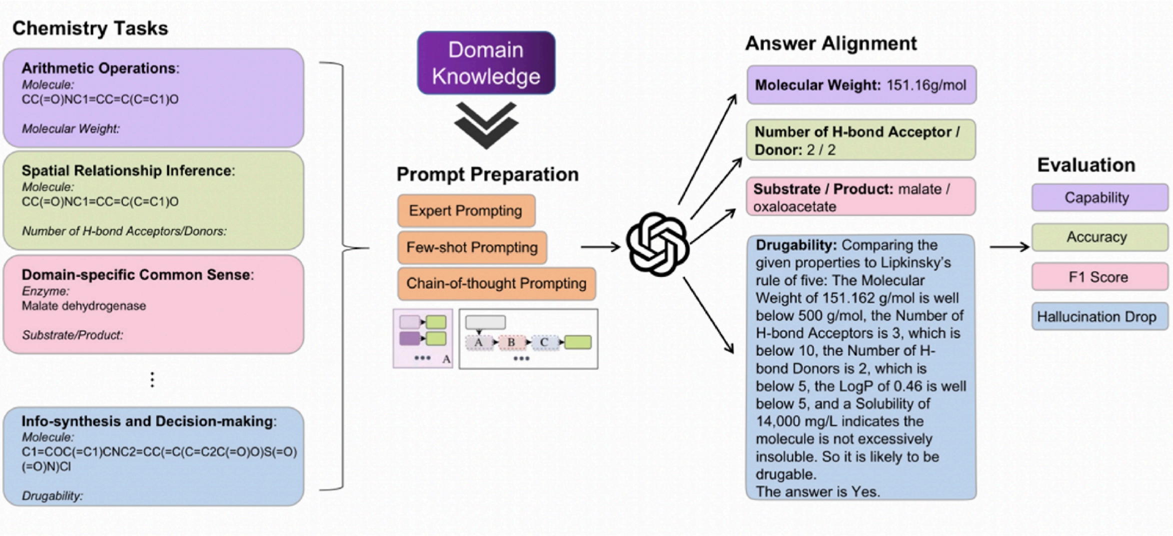

Chemistry Tasks
Arithmetic Operations:
Molecule:
CC(=O)NC1=CC=C(C=C1)O
Molecular Weight:
Spatial Relationship Inference:
Molecule:
CC(=O)NC1=CC=C(C=C1)O
Number of H-bond Acceptors/Donors:
Domain-specific Common Sense:
Enzyme:
Malate dehydrogenase
Substrate/Product:
Info-synthesis and Decision-making:
Molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Drugability:
Domain Knowledge
Prompt Preparation
Expert Prompting
Few-shot Prompting
Chain-of-thought Prompting
A
A
B
C
Answer Alignment
Molecular Weight: 151.16g/mol
Number of H-bond Acceptor / Donor: 2 / 2
Substrate / Product: malate / oxaloacetate
Drugability: Comparing the given properties to Lipkinsky's rule of five: The Molecular Weight of 151.162 g/mol is well below 500 g/mol, the Number of H-bond Acceptors is 3, which is below 10, the Number of H-bond Donors is 2, which is below 5, the LogP of 0.46 is well below 5, and a Solubility of 14,000 mg/L indicates the molecule is not excessively insoluble. So it is likely to be drugable.
The answer is Yes.
Evaluation
Capability
Accuracy
F1 Score
Hallucination Drop

Fig. 1. The Whole process of prompt engineering framework.

corresponding solutions (See Table 1) for the evaluation of LLM's capability, as described below. There are two most important criteria in molecule selection: practical applicability (for the purpose of assisting scientific application research) and the balance of various molecular types in the specific domain.

Organic Small Molecules: **40** molecules proven to have significant drug properties or potentials are selected and curated from Pubchem [52], each containing **9** crucial structural and physical-chemical properties.

Enzymes: **40** enzymes involved in significant metabolic pathways in vivo are selected and curated from UniProt database [53], each with **7** crucial sequence and functional information.

Crystal Materials: **40** representative crystals from some of the most critical structural and functional categories in materials science are collected from the Materials Project database [54], each with **16** crucial structural and energy properties.

A detailed enumeration and classification of all task types are contained in Appendix S.2. Due to the limitations in API callings of proprietary LLMs, it is very hard to test molecules on a larger scale (like for thousands of molecules), but we believe that the selected molecules are already very representative to demonstrate LLM's performance, and could pave the way for further applications in the future.

In evaluating the performance of LLM prompt engineering methods on different tasks, 4 significant metrics are introduced:

Capability: To measure LLM's capability to provide an answer for a certain task, regardless of its correctness. Its value takes 1 if the answer is effective otherwise 0.

Accuracy:To evaluate the extent to which LLM's answer is identical or close to the ground truth.

F1 Score: to measure LLM's predictive performance on multiple-choices questions, combining precision and recall. As a widely used metric in Statistics and Machine learning, F1 Score offers a more comprehensive evaluation compared to Accuracy, especially in cases of imbalance of precision and recall where accuracy might be high but does not reflect the true performance of LLM. A high F1 score is usually accompanied by high precision and recall, indicating stronger model performance.

Hallucination Drop: A metric put forward specifically for this work to quantify the discrepancy between an LLM's ability to answer questions (Capability) and the accuracy of those answers (Accuracy). It takes 1 minus the ratio of Accuracy and Capability as the value. This metric helps identify the hallucination level of LLM, as when the hallucinaton drop is high, the LLM tends to be trapped in generating hallucinated answers or scientific facts.

A detailed implementation of these metrics are listed in Appendix S.3.

In our approach, we utilize an LLM plugged-in automatic scheme to evaluate the metrics above. According to Table S.2 in Appendix, tasks can be divided into numerical and verbal ones, each of which takes a different manner to evaluate, respectively.

Numerical Tasks: All numerical tasks are transformed to the form of multiple-choices questions, as straightforward error estimation of the answer from ground truth can be strongly affected by unit and scale, and the form of multiple-choices makes it easier and more reasonable in evaluation across various tasks. Detailed implementation of tasks' transformation into multiple-choices questions are described in Appendix S.3. Metrics involved in numerical task evaluation are: **Capability, Accuracy, F1 Score and Hallucination Drop.**

While Capability, F1 Score and Hallucination Drop are evaluated in the normal form, the Accuracy of multiple choices questions is specifically defined. Full mark (1) is given if the option is exactly the ground truth. A partial score (0.4) is given if the value or range of the chosen option is adjacent to the ground truth. The complete scoring policy is listed in Appendix S.3.

Verbal Tasks: For verbal answers, the LLM is guided by a series of grading examples coordinated to the specific question types and then required to give a grade to an answer. Detailed prompts for LLM's grading tasks are listed in Appendix S.4. Metrics involved in verbal task evaluation are: **Capability, Accuracy and Hallucination Drop.**

While Capability and Hallucination Drop are evaluated in the normal form, the Accuracy of verbal tasks take discrete values among {0, 0.2, 0.4, 0.6, 0.8, 1}. Score 0 means that the answer is completely irrelevant to the ground truth, while the scores {0.2, 0.4, 0.6, 0.8} imply part of the answer aligned with the ground truth, extent to which increases with the value. Score 1 corresponds to answers intrinsically the same as ground truth.

It is worth emphasizing that we believe the LLM plugged-in automatic scheme above for evaluation could bear skepticism on fairness and effectiveness, as LLM's evaluation process is independent from LLM's predictive task performing in the last step, implying the LLM would not take past memories of task performing or "know" the answers were generated by itself, and thus is unlikely to "cheat" on the grading process.

Fig. 2 shows the flow chart of question construction and answer evaluation process. Data from 3 material categories are extracted and combined to form proper questions (some are in the form of multiple-choices questions). When the raw answers are acquired, they need to be checked for validity, and then aligned to proper answer forms. Ultimately the answers are automatically graded.

2.2. *Scientific prediction as a LLM question answering problem*

In the LLM era, scientific prediction can be considered as a question answering task leveraging the zero-shot/few-shot reasoning power of LLM. It is demonstrated that by providing in-context hints to language model with size large enough for emergence to happen, the model can excavate knowledge learned from pre-trained data and well-perform the question answering task [55]. As an approach to enhance LLM's capability on specific domains or tasks, prompt engineering significantly reduces the need for extensive task-specific datasets as required in LLM fine-tuning paradigm, making it an effective in-context learning method for LLM enhancement.

The process of prompt engineering could be mathematically formalized [56]. Let Q be the question, P be the prompt, A be the answer by LLM, prompt engineering process is to determine the context of prompt words P such that the answer A could be given effectively by LLM:

$$A = f(P, Q) \tag{1}$$

where f is the LLM.

A prompt optimization objective is to find:

$$\arg\ \max_P g(f(P, Q), S) \tag{2}$$

where S is the ground truth solution, and g is a evaluation function which measures how much the LLM answer A is in accordance with the ground truth solution S.

For our dataset $D = \{Q_i, S_i\}_i^n$, the general prompt optimization objective is to find the P that maximizes the expectation over the dataset:

$$\arg\ \max_P \mathbb{E}_{Q,S\in D}\, g(f(P, Q), \mathrm{S}) \tag{3}$$

2.3. *Common prompt engineering techniques and domain-knowledge embedded prompt engineering*

The essence of prompt engineering is to harness the full potential of LLMs in diverse applications by ensuring they respond in a manner that is most aligned with the user's intent and the task at hand. We give a brief introduction to several mainstream prompt engineering methods:

Zero-shot Prompting: Zero-shot Prompting requires LLM to answer the given question directly without providing any data or example questions in the context (See Fig. 3 (a)).

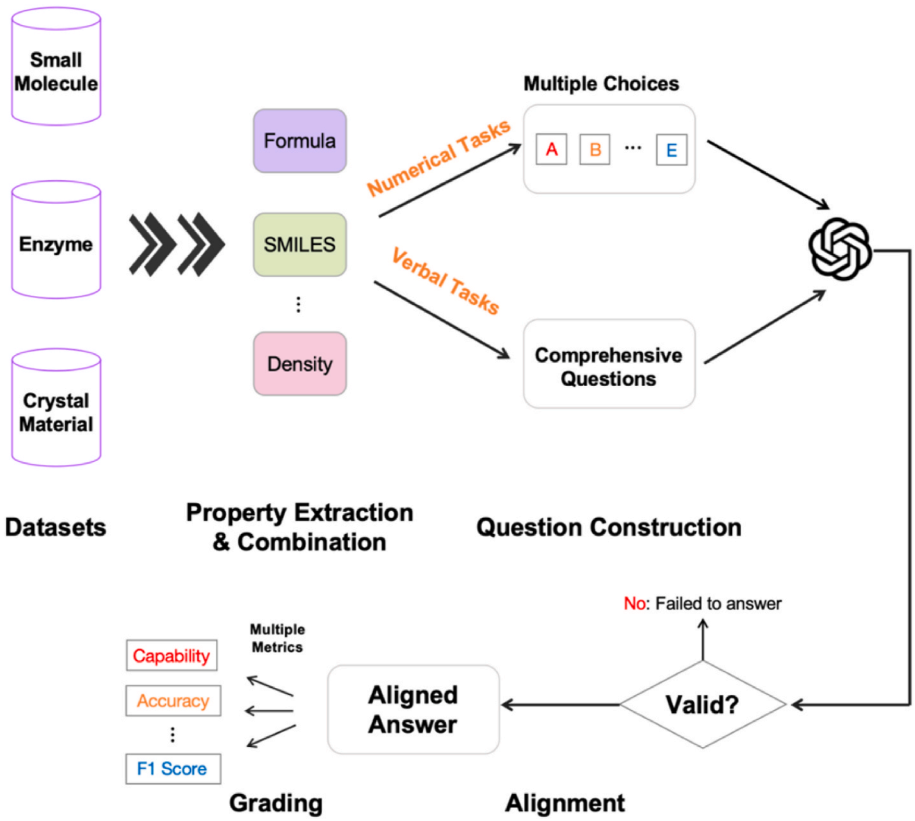
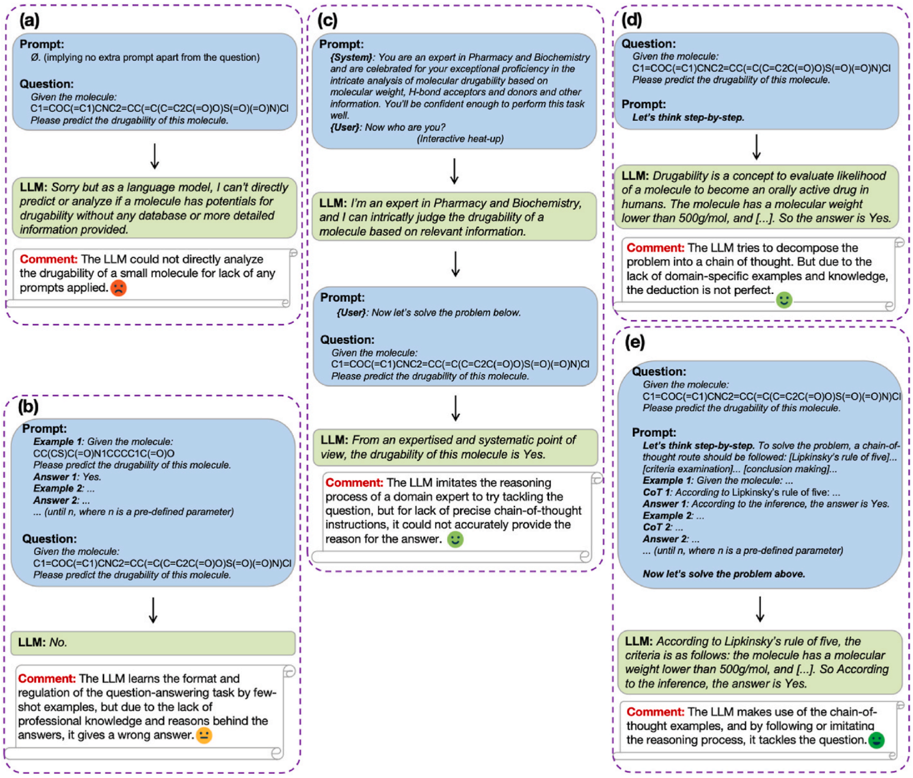


Fig. 2. Question construction, answer alignment and grading process.



(a)

Prompt:
Ø. (implying no extra prompt apart from the question)

Question:
Given the molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Please predict the drugability of this molecule.

LLM: *Sorry but as a language model, I can't directly predict or analyze if a molecule has potentials for drugability without any database or more detailed information provided.*

Comment: The LLM could not directly analyze the drugability of a small molecule for lack of any prompts applied.

(b)

Prompt:
***Example 1**: Given the molecule:*
CC(CS)C(=O)N1CCCC1C(=O)O
Please predict the drugability of this molecule.
***Answer 1**: Yes.*
***Example 2**: ...*
***Answer 2**: ...*
... (until n, where n is a pre-defined parameter)

Question:
Given the molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Please predict the drugability of this molecule.

LLM: *No.*

Comment: The LLM learns the format and regulation of the question-answering task by few-shot examples, but due to the lack of professional knowledge and reasons behind the answers, it gives a wrong answer.

(c)

Prompt:
***{System}**: You are an expert in Pharmacy and Biochemistry and are celebrated for your exceptional proficiency in the intricate analysis of molecular drugability based on molecular weight, H-bond acceptors and donors and other information. You'll be confident enough to perform this task well.*
***{User}**: Now who are you?*
(Interactive heat-up)

LLM: *I'm an expert in Pharmacy and Biochemistry, and I can intricatly judge the drugability of a molecule based on relevant information.*

Prompt:
***{User}**: Now let's solve the problem below.*

Question:
Given the molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Please predict the drugability of this molecule.

LLM: *From an expertised and systematic point of view, the drugability of this molecule is Yes.*

Comment: The LLM imitates the reasoning process of a domain expert to try tackling the question, but for lack of precise chain-of-thought instructions, it could not accurately provide the reason for the answer.

(d)

Question:
Given the molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Please predict the drugability of this molecule.

Prompt:
Let's think step-by-step.

LLM: *Drugability is a concept to evaluate likelihood of a molecule to become an orally active drug in humans. The molecule has a molecular weight lower than 500g/mol, and [...]. So the answer is Yes.*

Comment: The LLM tries to decompose the problem into a chain of thought. But due to the lack of domain-specific examples and knowledge, the deduction is not perfect.

(e)

Question:
Given the molecule:
C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl
Please predict the drugability of this molecule.

Prompt:
***Let's think step-by-step.** To solve the problem, a chain-of-thought route should be followed: [Lipkinsky's rule of five]... [criteria examination]... [conclusion making]...*
***Example 1**: Given the molecule: ...*
***CoT 1**: According to* Lipkinsky's rule of five: ...
***Answer 1**: According to the inference, the answer is Yes.*
***Example 2**: ...*
***CoT 2**: ...*
***Answer 2**: ...*
... (until n, where n is a pre-defined parameter)

Now let's solve the problem above.

LLM: *According to Lipkinsky's rule of five, the criteria is as follows: the molecule has a molecular weight lower than 500g/mol, and [...]. So According to the inference, the answer is Yes.*

Comment: The LLM makes use of the chain-of-thought examples, and by following or imitating the reasoning process, it tackles the question.

Fig. 3. Mainstream prompt engineering methods Illustration
(a) zero-shot prompting; (b) few-shot prompting; (c) expert prompting; (d) zero-shot CoT prompting; (e) few-shot CoT prompting.

Few-shot Prompting: In Few-shot Prompting, an LLM is presented with several *demonstrations*, (i.e. question-answer pairs within the prompt context), better equipping the LLM to understand and replicate the response format and content. The demonstrations in prompt can be formalized to:

$$P = \{(Q_1, A_1), \ldots, (Q_m, A_m)\} \tag{4}$$

where *m* is the number of examples [55]. (See Fig. 3 (b))

Expert Prompting: Role-play instructions have demonstrated their effectiveness in harnessing the potential of LLMs [56]. By guiding LLMs step-by-step into assuming the role of domain experts, they can generate responses akin to those written by experts (See Fig. 3 (c)).

Zero-shot CoT (Chain-of-Thought) Prompting: By eliciting a sequential, step-by-step reasoning process to effectively address complex tasks, CoT enables the model to break down a task into its constituent parts, offering a clear and logical pathway to the solution [36]. In particular, Zero-shot CoT prompting involves adding *"Let's think step by step"* to the prompt as a trigger-sentence (See Fig. 3 (d)).

Few-shot CoT Prompting: In addition to adding *"Let's think step by step"* to the prompt like Zero-shot CoT, Few-shot CoT provides several examples of Thought-Chain in solving similar problems to assist LLM perform the current task in a similar manner [57]. The *demonstrations* in prompt (See Fig. 3 (e)) can be formalized to:

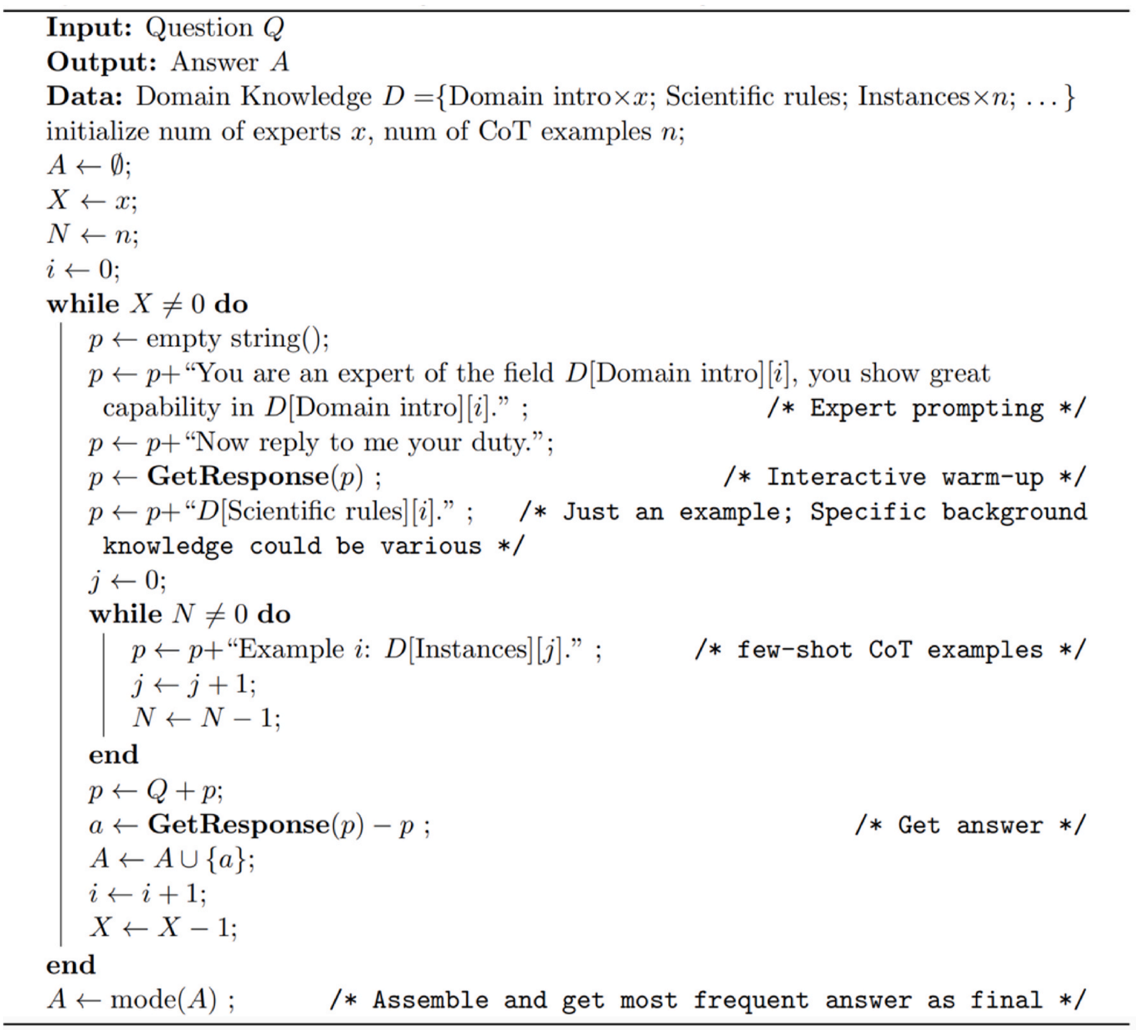
$$P = \{(Q_1, C_1, A_1), \ldots, (Q_m, C_m, A_m)\}. \tag{5}$$

A significant limitation of these prompt engineering methods is that they do not incorporate domain expertise as guidance for problem-solving, which considerably restricts the capabilities of LLMs in numerous domain-specific tasks. Moreover, since addressing many domain-specific challenges involves intricate cognitive processes, it is imperative to strategically combine various prompt engineering techniques at different stages to achieve optimality.

Here we propose a domain-knowledge embedded prompt engineering strategy that integrates chemistry knowledge into language model. The prompting scheme takes the form of multi-expert mixture. Each expert takes part in role playing and are given a few shots of CoT demonstrations integrated with expertise domain knowledge or instructions.

Here, incorporating *domain knowledge* essentially involves integrating the thought processes of chemistry/materials experts. This contrasts with the conventional zero-shot CoT approach, which merely prompts LLMs to engage in a chain of thought. By doing so, it offers more precise background knowledge and exemplifies more accurate human reasoning. In Algorithm 1 and Algorithm 2, we illustrate how Domain-knowledge embedded prompts are typically structured.

Algorithm 1. Domain-knowledge embedded prompting



```
Input: Question $Q$
Output: Answer $A$
Data: Domain Knowledge $D$ ={Domain intro$\times x$; Scientific rules; Instances$\times n$; ...}
initialize num of experts $x$, num of CoT examples $n$;
$A \leftarrow \emptyset$;
$X \leftarrow x$;
$N \leftarrow n$;
$i \leftarrow 0$;
while $X \neq 0$ do
    $p \leftarrow$ empty string();
    $p \leftarrow p+$"You are an expert of the field $D$[Domain intro][$i$], you show great
     capability in $D$[Domain intro][$i$]." ;                  /* Expert prompting */
    $p \leftarrow p+$"Now reply to me your duty.";
    $p \leftarrow$ GetResponse($p$) ;                          /* Interactive warm-up */
    $p \leftarrow p+$"$D$[Scientific rules][$i$]." ;   /* Just an example; Specific background
     knowledge could be various */
    $j \leftarrow 0$;
    while $N \neq 0$ do
        $p \leftarrow p+$"Example $i$: $D$[Instances][$j$]." ;       /* few-shot CoT examples */
        $j \leftarrow j+1$;
        $N \leftarrow N-1$;
    end
    $p \leftarrow Q+p$;
    $a \leftarrow$ GetResponse($p$) $- p$ ;                                /* Get answer */
    $A \leftarrow A \cup \{a\}$;
    $i \leftarrow i+1$;
    $X \leftarrow X-1$;
end
$A \leftarrow$ mode($A$) ;        /* Assemble and get most frequent answer as final */
```

Algorithm 2. GetResponse Function

```
Input: Prompt $p$
Output: Response $r$
$r \leftarrow$ empty string();
$r \leftarrow$ call to LLM API service with $p$;
return $p + r$ ;   /* Record both question and answer as complete history */
```

The domain-knowledge embedded prompting first involves leveraging multiple role-playing experts, to receive background knowledge and adapt into the role. These experts apply domain-specific problem-solving capability through few-shot CoT examples, which utilizes more detailed instructions and knowledge to break down the problem into smaller steps. Finally, the outputs from all experts would be assembled through the principle of "minority submission to the majority".

The full documentation of all domain-knowledge prompts are listed in Appendix S.4. The high-level schemes of these strategies are delineated in Figs. 3 and 4.

In the following chapters, we compare this prompt engineering method proposed above to other generic prompt engineering methods including zero-shot prompting, few-shot prompting, expert prompting, and CoT prompting.

3. Results

In this section, we first present the overall benchmarks of prompt engineering methods over all tasks. Then we make detailed comparisons over different task types, CoT complexities and material types. In the last section, 3 case studies on representative molecules are conducted using our tailored domain-knowledge embedded prompt engineering method to illustrate the effectiveness of prompt engineering in assisting crucial scientific research topics.

3.1. *Summary of overall performance*

In our study, we evaluated 5 different prompt engineering strategies across three datasets (small molecule, enzyme, and crystal material), each yielding 3 sets of answers for robustness. The LLM model being evaluated is ‘gpt-3.5-turbo-1106’ [58] through official API calling. The prompt engineering strategies included zero-shot, few-shot, expert, and zero-shot CoT, along with domain-knowledge expert CoT (ours).

The overall evaluation results on 3 datasets are shown in Figs. 5 and 6.

Our domain-knowledge embedded prompt engineering method outperforms other conventional prompt engineering techniques on most tasks and metrics. In nearly all tasks on enzymes and crystal materials, and more than 50 % of the tasks on small molecules, our method’s performance is very significantly higher than other methods, while on tasks: Molecular Density, Molecular Weight, Number of Amino Acids and Active Sites, our method does not demonstrate obvious advantages.

In the following sections, we make more detailed comparisons for different tasks and molecules. Due to space limitation, we only present the key findings in the following sections. In Section 3.2, we compare these method’s performance on different task types, while in Section 3.3, we delve into the correlation between prompt engineering method’ performance and CoT complexity. Finally we compare prompt engineering methods’ effectiveness on different types of materials.

3.2. *Comparison by task types*

In this section, we compare various prompt engineering methods performance on different types of tasks, and the detailed classifications

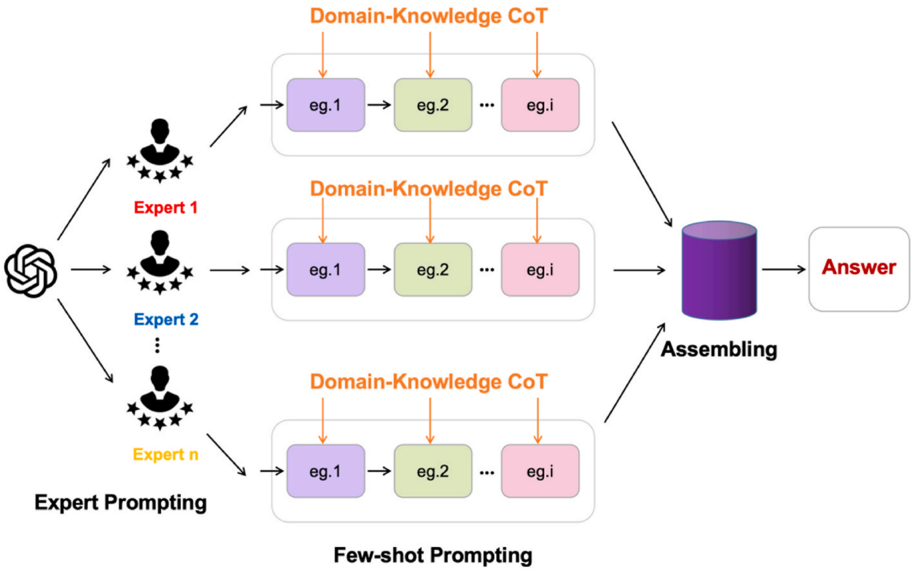


Fig. 4. The Whole Process of Domain-Knowledge Prompt Engineering Method
(N experts are assembled to give answers separately, and “eg.i” represents few-shot examples integrated with CoT knowledge).

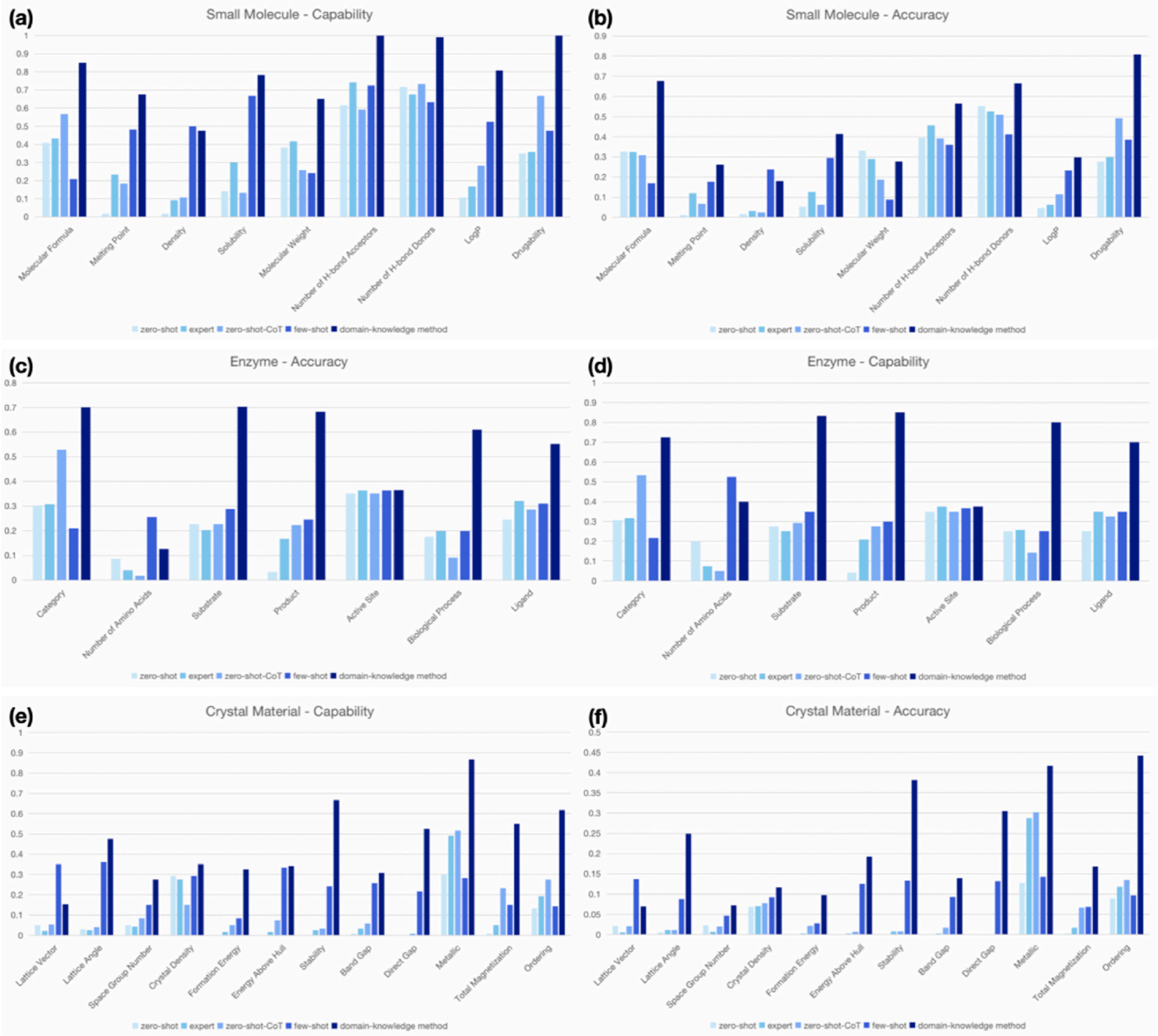


Fig. 5. Capability and Accuracy for All Tasks
(a) Capability on Small Molecules; (b) Accuracy on Small Molecules; (c) Capability on Enzymes; (d) Accuracy on Enzymes; (e) Capability on Crystal Materials; (f) Accuracy on Crystal Materials
(The abscissa is each task under each data set; the ordinate is the metric scores of different prompt engineering methods under the task.).

can be referred to in Table S2 and S.3 in supplementary materials. Each task type presents unique challenges and necessitates different inferencing abilities from the LLM. After aggregation, the performances of 5 prompt engineering methods on different question groups are shown in Fig. 7.

(1) **Domain-knowledge embedded prompt engineering method outperforms traditional prompt engineering methods on all question types.** Through a comprehensive evaluation across various groups of prediction tasks, focusing on four crucial indices - "Capability", "Accuracy", "F1 Score" and "Hallucination Drop", our domain-knowledge embedded prompt engineering method consistently outperforms traditional prompt engineering strategies. This superiority is evident in the substantial enhancement of both capability and accuracy metrics, with the most notable improvements exceeding a 100 % boost. Such findings unequivocally demonstrate that integrating domain-specific knowledge into prompt engineering substantially elevates the effectiveness of generic prompt engineering techniques.

(2) **LLM performs better for answers derived from logical reasoning than answers based on experimental data.** This tendency is further amplified in our domain-specific prompt engineering method, where a more tailored prompt engineering strategy is applied. As shown in Fig. 7(a) and (b) and (c), it consistently leads to more significant improvements in tasks involving logical deduction compared to other prompt engineering methods. This disparity in performance can be attributed to the fact that LLMs, with refined prompt engineering, can engage in a sophisticated Chain-of-Thought process, enabling LLMs to excel in tasks that demand intricate reasoning and problem-solving skills. However, despite being trained on various scientific databases, LLMs do not excel in precisely replicating exact data values. This brings about their ability to process and reason through information well rather than serve as direct conduits for data retrieval.

(3) **LLM performs better on verbal tasks compared to numerical tasks.** When faced with tasks that require a numerical response, (actually in formats involving multiple choices), LLMs tend to exhibit weaker performance. This is evident in both capability and accuracy metrics across various prompt engineering methods, with numerical answers derived from experimental data showing the least favorable results (Fig. 7 (a), (b), (c)). When LLMs engage in numerical reasoning, their capability scores are notably higher (Fig. 7 (a)), but this advantage is tempered by

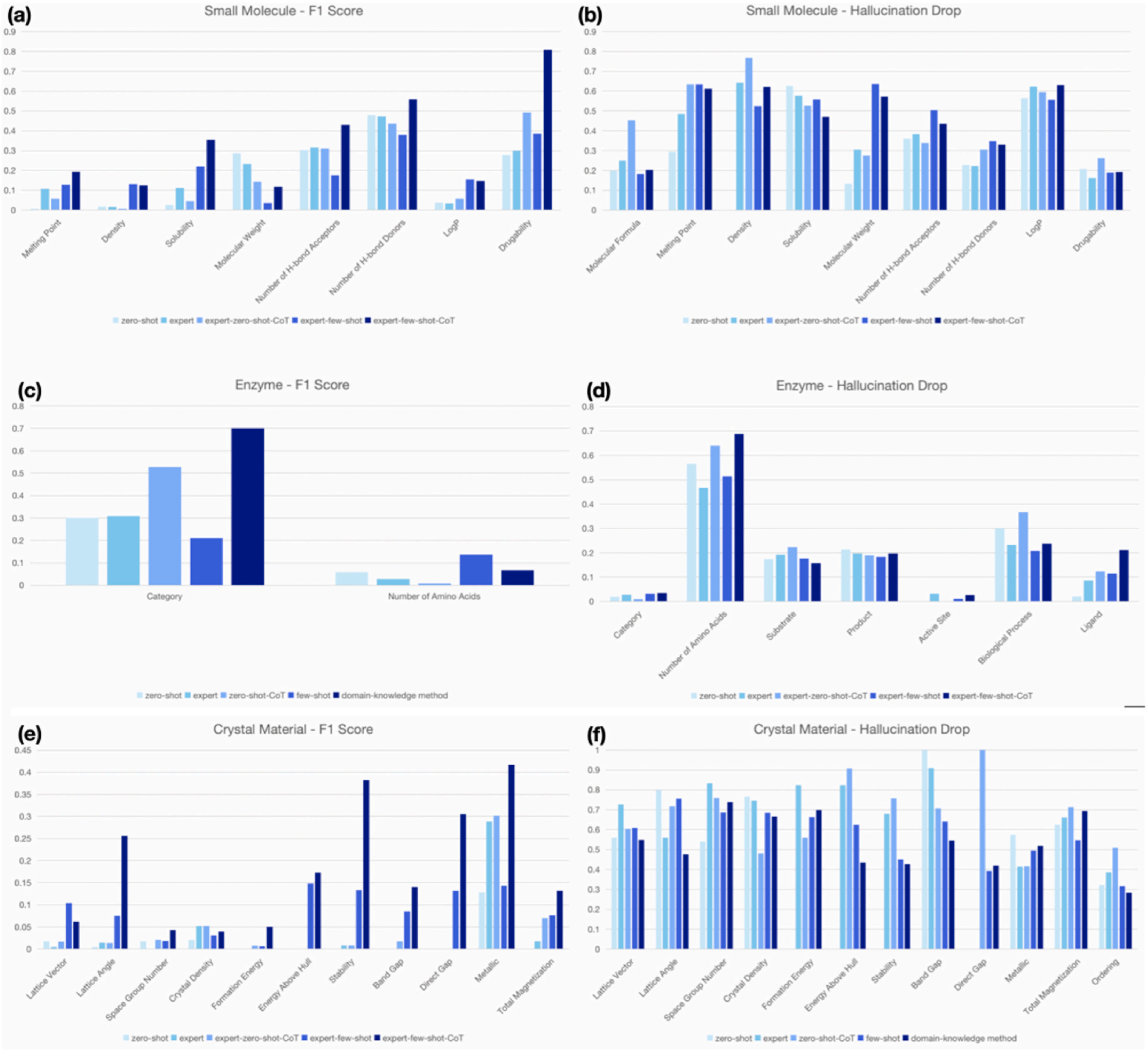


Fig. 6. F1 Score and Hallucination Drop for All Tasks
(a) F1 Score on Small Molecules; (b) Hallucination Drop on Small Molecules; (c) F1 Score on Enzymes; (d) Hallucination Drop on Enzymes; (e) F1 Score on Crystal Materials; (f) Hallucination Drop on Crystal Materials
(The abscissa is each task under each data set; the ordinate is the metric scores of different prompt engineering methods under the task.).

significant issues with hallucinations, which adversely affect the accuracy of these responses. In fact, even when the most advanced prompt engineering methods are applied, the accuracy of logical numerical answers is surpassed by that of logical answers. This trend underscores a recognized weakness of LLMs in number-related tasks, as evidenced by several research studies [59,60].

(4) **Domain-knowledge embedded prompt engineering method effectively reduces hallucination.** The metric of hallucination drop serves as a barometer for the average quality of answers produced by LLMs under different prompt engineering strategies. As shown in Fig. 7 (d), the question type of numerical answer by logic is the only category where an increase in hallucination is observed as the domain-knowledge embedded prompt engineering is applied. In the other three question types, the incorporation of domain-specific knowledge into the prompt engineering process effectively curtails the occurrence of hallucinations. Notably, the question types 'numerical answer by experimental data' and 'verbal answer by logic' emerge as frontrunners, registering the top two lowest scores in hallucination drop. This outcome underscores the precision and effectiveness of domain-knowledge embedded prompt engineering methods in enhancing the reliability and accuracy of LLM responses.

The results from a more detailed classification based on reasoning paradigm also draw some intriguing conclusions below, showing the distinctive strengths and drawbacks in LLM reasoning.

(5) **LLM performs poorly on arithmetic tasks.** These tasks revolving around basic counting, adding, and multiplying abilities, ostensibly require less sophisticated cognitive skills compared to tasks that necessitate spatial imagination or intense domain-knowledge based reasoning, but the performance of LLMs in these arithmetic tasks is unexpectedly subpar. Despite scoring high in capability, LLMs do not exhibit a corresponding lead in accuracy, showing higher occurrence of hallucinations in these tasks. In fact, the accuracy of LLMs in arithmetic tasks is not only significantly outpaced by domain knowledge literal reasoning tasks but also closely rivalled by spatial relationship tasks (Fig. 8 (b)). Notably, even the application of CoT heuristics in the reasoning process does not substantially mitigate this issue. This is evident in the Hallucination Drop metric, where both zero-

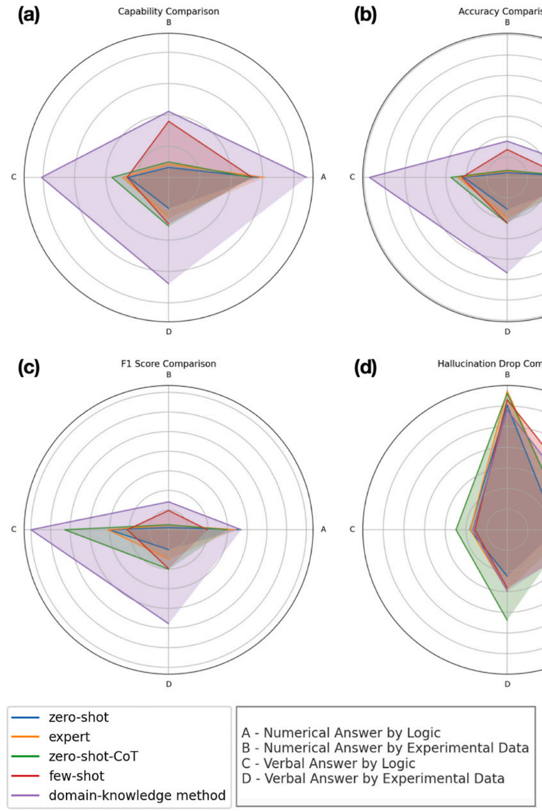


Fig. 7. Prompt engineering performances by output Type
(a) capability on different tasks; (b) accuracy on different tasks; (c) F1 score on different tasks; (d) hallucination drop on different tasks.

shot-CoT and domain-knowledge embedded method exhibit a higher incidence of hallucination phenomena in arithmetic tasks compared to others (Fig. 8 (d)).

(6) **LLM is incapable on many information retrieval tasks.** These tasks, which cannot be effectively addressed through reasoning alone, generally exhibit poorer performance compared to those based purely on reasoning. As depicted in Fig. 8(a) and (b) and (c), tasks involving the retrieval of both common and uncommon properties record the lowest capability and accuracy scores, with tasks involving uncommon information faring slightly worse. Prompt engineering falls short in information retrieval, primarily due to its inability to provide direct access to external databases, but a well-crafted, domain-specific prompt can still marginally improve LLM performance by encouraging a more detailed response, as indicated by the higher capability scores for domain-knowledge prompt engineering. Despite this, the challenge of mitigating hallucinations remains formidable, with the highest incidence of hallucination observed in these types of tasks when using domain-knowledge prompts.

(7) **Verbal reasoning tasks get largest boosting with domain-knowledge embedded prompt engineering method.** In 5 question types classified by reasoning paradigms, "Domain Knowledge Literal Reasoning Tasks" distinctly stand out, especially when enhanced by domain-knowledge embedded prompt engineering methods. This category of tasks not only achieves the highest capability and accuracy scores overall but also maintains a relatively low level of hallucinations.This demonstrates well-crafted prompts can, in a remarkably efficient manner, stimulate the latent capabilities of LLMs, enabling them to generate answers with heightened confidence and precision.

3.3. Comparison by CoT complexity

In this section, we compare different prompt engineering methods' distinction under a variety of CoT complexities, in order to depict our tailored prompt engineering method's superiority under different CoT complexities. We propose that the quantity of additional properties added in CoT prompts serves as a viable metric for gauging the complexity of the CoT process. This metric reflects the extent of extra information that is integrated into the CoT reasoning, which in turn influences the complexity and depth of the reasoning required. To operationalize this, we have categorized tasks based on the number of additional properties provided in each question, as shown on Table S.4 in Appendix.

It is, however, worth noticing that the number of additional properties provided (namely, the complexity of CoT) does not necessarily correlates to the difficulty of questions. The aggregated results are shown in Fig. 9.

(1) **Domain-Knowledge Embedded Prompt Engineering produces greatest performance lift in tasks with most complicated CoT formulation.** In scenarios where LLMs are presented with different amounts of additional information for task execution, the domain-knowledge embedded prompt engineering method emerges as the most effective, outshining others in three key performance metrics: "Capability", "Accuracy", and "F1 Score". Specifically, it excels remarkably in "Tasks with Multiple Additional Properties" (Fig. 9 (a), (b), (c)). This highlights the advantage of domain-knowledge prompts in enhancing LLM performance in tasks that demand a complex CoT formulation. Furthermore, even for simpler zero-shot CoT method, this benefit makes it reverse the lead of few-shot method in "Tasks with Multiple Additional Properties", especially on F1 Score (Fig. 9 (c)). This aligns well with the intuitive understanding of CoT in enhancing inference-related capabilities.

(2) **In-Context Information Could Effectively Reduce Hallucination Level.** Tasks supplied with the most in-context extra information consistently exhibit the lowest levels of hallucination across all prompt engineering methods, as shown in Fig. 9 (d). This trend holds true regardless of whether the prompt engineering method incorporates domain-knowledge features. A notable observation is that many tasks in the "Tasks with Multiple Additional Properties" category are inherently complex and challenging. For example, predicting the drugability of a small molecule often necessitates a thorough and intricate examination under Lipinski's Rule of Five. Similarly, calculating the crystal density of a substance involves complex computations, including the determination of relative molecular mass of a unit cell, the measurement of unit cell volume, and intricate unit transformations. The surprisingly low hallucination levels is indicative of the effectiveness of providing additional in-context information, suggests that enriching LLM prompts with more contextual information and factual details may substantially enhance the robustness and reliability of the generated content.

3.4. Comparison by material differences

In this section, a detailed comparison of prompt engineering accuracy on three types of materials will be portrayed. For clarity, we only focus on our tailored prompting method (namely the domain-knowledge embedded prompting)'s performance on small molecules, enzymes and crystal materials with divergent material traits. The methodology employed to quantify the differences among these materials will be elaborated upon in the following paragraphs.

For small molecules, we propose two indicators—**molecular weight** and **elemental composition**—to differentiate the complexity of various molecules. This is predicated on the rationale that more complex

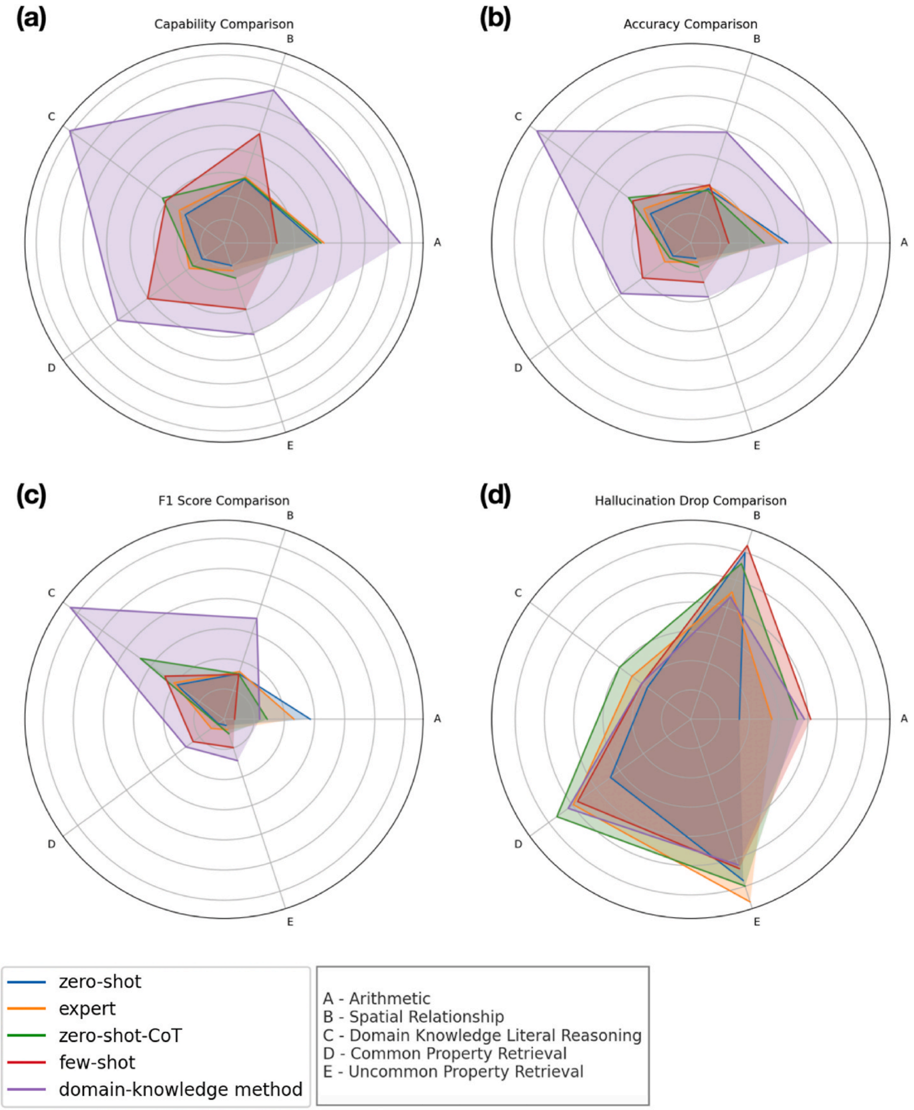


Fig. 8. Prompt engineering performances by reasoning Paradigm
(a) capability on different tasks; (b) accuracy on different tasks; (c) F1 score on different tasks; (d) hallucination drop on different tasks.

molecules typically necessitate a higher level of analytical effort, which could inversely affect accuracy. We aim to investigate whether this assumption aligns with the empirical results obtained from our study.

For enzymes, we also employ two indicators to discern the difficulty in predicting properties among different enzymes: **enzyme scale**, quantified by the number of amino acids, and the **current depth of research**, denoted as the number of reviewed publications recorded on Uniprot [53]. It is hypothesized that more complex enzymes, characterized by longer peptide chains and a lesser degree of comprehensive research, necessitate a higher analytical effort, potentially reducing accuracy. We intend to explore whether this hypothesis is consistent with the overall findings of our analysis.

For crystalline materials, we utilize two indicators to gauge complexity: **formula complexity**, which reflects the complexity of a single unit cell, and **unit cell symmetry**, denoted by the crystal system to which it belongs. The underlying premise is that more complex crystalline materials demand a more substantial analytical effort, which could, in turn, diminish accuracy. We will investigate whether this premise aligns with the collective results of our study.

(1) **The prediction accuracy of LLMs deteriorates for larger and more complex organic molecules.** As the molecular weight increases and the elemental composition becomes more diverse, we observe a gradual decline in the LLM's prediction accuracy. Specifically, molecules comprising more than five distinct elements exhibit significantly poorer performance compared to those with fewer components. Moreover, when the molecular weight exceeds 300 g/mol, the overall accuracy for single molecule predictions generally falls below 30 %, as shown in Fig. 10 (a). Furthermore, large organic molecules are less commonly found in literature compared to smaller molecules, exacerbating the difficulty of LLM's information retrieval.

(2) **The accuracy of LLMs in predicting properties of specific enzymes aligns closely with the depth of current research on these enzymes but shows a weak correlation with the enzymes' size.** The number of reviewed publications recorded on Uniprot, which signifies the academic community's past research focus on an enzyme, demonstrates a strong correlation with the LLM's prediction performance. The more thoroughly an enzyme is researched and understood, the higher the accuracy of LLM predictions. Most enzymes with low prediction accuracy concentrate in areas with low number of reviewed publications, as shown in Fig. 10 (b). However, there appears to be no explicit relationship between the size of the enzyme, measured by the

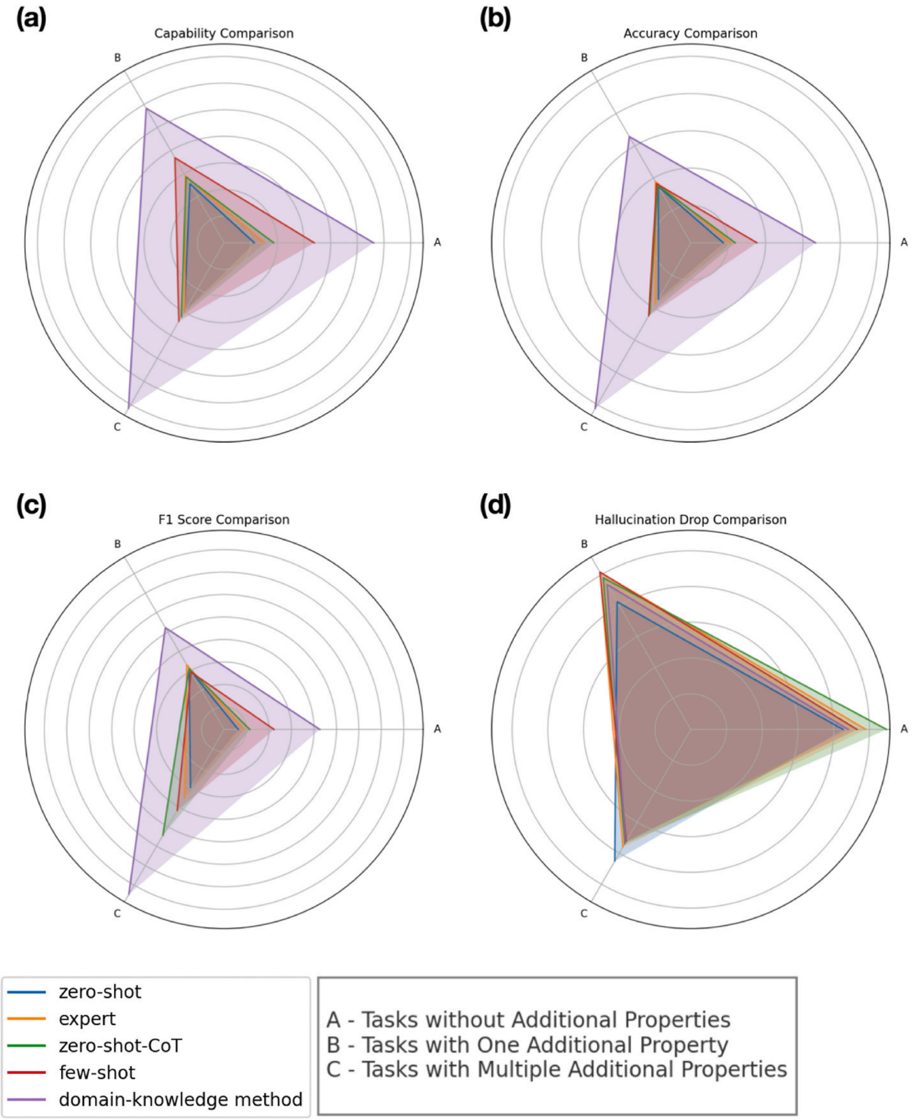


Fig. 9. Prompt engineering performances by CoT Complexity
(a) capability on different tasks; (b) accuracy on different tasks; (c) F1 score on different tasks; (d) hallucination drop on different tasks.

number of amino acids, and the accuracy of LLM predictions. This outcome suggests that the predictive ability of LLMs for enzymes primarily relies on information retrieval, specifically from scientific literature reports, rather than on the direct analysis of the enzyme's structure.

(3) **The prediction accuracy of LLMs decreases for crystalline materials with larger, more complex compositions.** As the prediction target's gauge complexity increases, indicated by formula complexity, there is a gradual decline in the LLM's prediction accuracy. Crystals comprising more than four elements perform significantly worse than those with fewer components. Additionally, when the number of formula atoms exceeds 10, the overall accuracy for single crystal predictions generally falls below 15 %, as shown in Fig. 10 (c). Apart from the intrinsic complexity of crystals to bring difficulty in analysis, since most prediction tasks for crystalline materials in our datasets do not require inference and mainly rely on data retrieval, the rarity of large crystals in the literature compared to more common crystalline materials increases the difficulty of LLM's information retrieval.

(4) **The prediction accuracy of LLMs concerning crystalline materials demonstrates a notable correlation with unit cell symmetry. Specifically, crystals belonging to the Trigonal, Cubic, or Hexagonal lattice systems are more likely to yield better predictions.** The reason for this is twofold: first, these structures are inherently more regular and defined, making them easier subjects for inferential analysis. Secondly, these types of crystal structures are more readily studied and characterized by modern crystallography instruments and techniques, such as X-ray diffraction and electron microscopy, leading to a richer presence in scientific literature. This abundance of data enhances the LLM's ability to retrieve relevant information, thereby improving prediction accuracy for crystals with these symmetries.

In conclusion, these empirical evidences presented supports **the intuitive notion that domain-knowledge embedded prompts enhance the performance of LLMs to different extents.** Firstly, the prompts' inferential capabilities are closely tied to the complexity of the analytical subject matter. Secondly, their proficiency in retrieval is correlated with the depth of contemporary academic research, suggesting that well-crafted prompts can effectively mine the latent knowledge absorbed during the LLM's pre-training phase.

Ultimately, these findings pose future challenges for leveraging LLMs

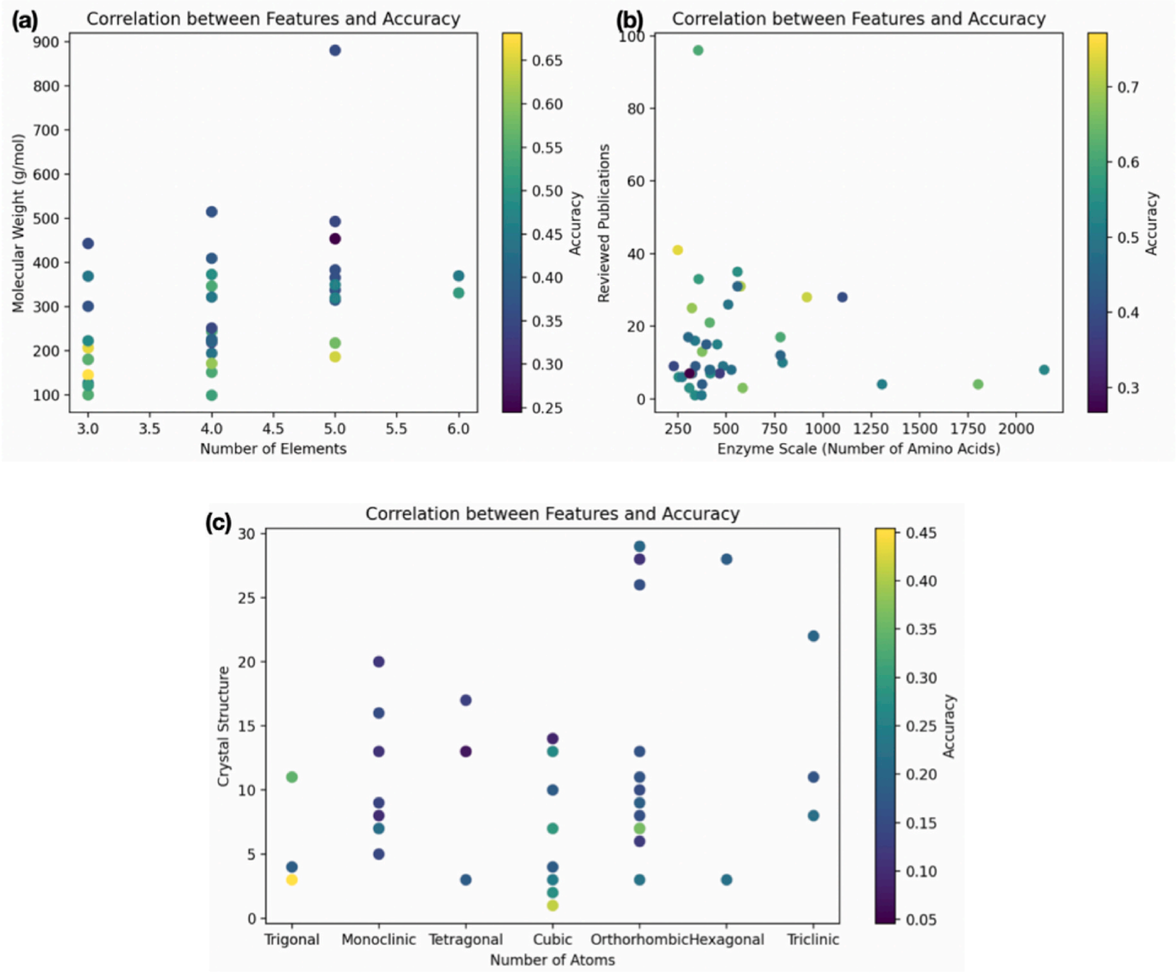


Fig. 10. Prompt engineering performances on different Materials
(a) accuracy distribution on small molecules; (b) accuracy distribution on enzymes; (c) accuracy distribution on crystal materials.

to aid scientific inquiries into complex and novel molecules that are rarely encountered or underrepresented in the academic literature. By addressing these challenges, LLMs could potentially revolutionize the approach to research in the synthesis, analysis, and applications of such molecules, thereby expanding the frontiers of scientific knowledge.

3.5. Case studies

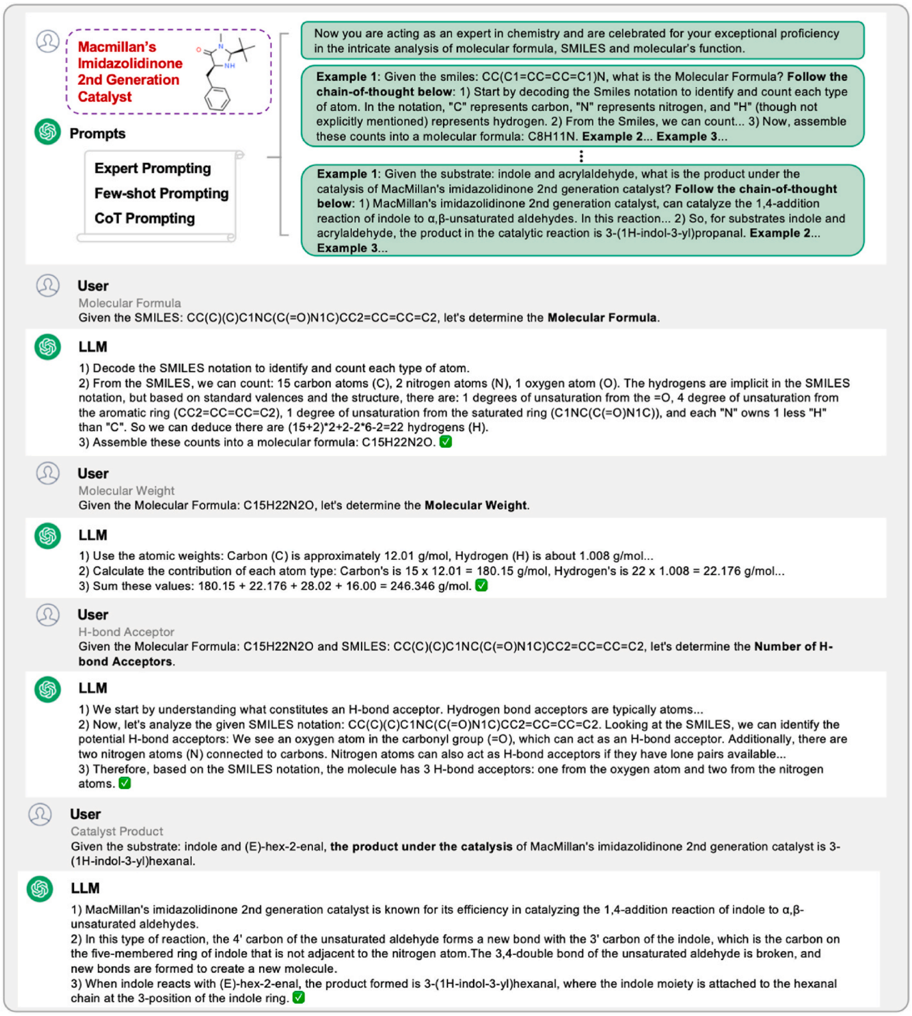
To elucidate the efficacy of the domain-knowledge embedded prompt engineering method in addressing highly domain-specific tasks, we have meticulously designed three case studies. These studies centerpiece the investigation of three materials of profound chemical importance, both in terms of academic research and industrial applications, utilizing our bespoke prompt engineering method that incorporates chemistry-specific domain knowledge. To enhance clarity and conciseness, we illustrate a single expert's prompt engineering workflow, omitting the assembly of contributions from multiple experts, as this singular demonstration already effectively showcases how our prompt engineering method significantly impacts the performance of the LLM.

In the first case study, we direct our attention to the **MacMillan's imidazolidinone 2nd generation catalyst**, (2*S*, 5S)-(−)-2-*tert*-Butyl-3-methyl-5-benzyl-4-imidazolidinone. The MacMillan catalyst, a groundbreaking advancement in the field of chemistry, was distinguished by the Nobel Prize in Chemistry in 2021 for its seminal contributions to the development of organocatalysis [61,62]. This innovation has had a transformative impact on both synthetic chemistry and the broader chemical industry, enabling more efficient and environmentally friendly catalytic processes that are pivotal in the synthesis of complex molecules.

The first case study aims to assess the capability of LLMs in assimilating the intricate details of this molecule and in delineating its potential applications. By employing our domain-knowledge embedded prompt engineering method, we seek to uncover how LLMs can be leveraged to provide insights into the reactivity, selectivity, and scope of application of the MacMillan catalyst, thereby enhancing the efficiency and productivity of chemical research in this area.

As shown in Fig. 11, by utilizing our tailored prompts, the LLM effectively elucidated the fundamental attributes of MacMillan's second-generation imidazolidinone catalyst, demonstrating its proficiency in the analysis of SMILES sequences and elementary arithmetic operations. Additionally, armed with the catalyst's mechanism and illustrative examples, the LLM was able to accurately anticipate the catalytic products from specified substrates, thereby highlighting the model's capacity to inform and potentially guide practical and industrial applications of catalysts. In light of this case study, it is evident that LLMs, embedded with domain-knowledge prompts, have the potential to significantly facilitate the development and optimization of catalysts for chemical reactions, thereby enhancing the efficiency and selectivity of synthetic processes in the field of chemistry. The complete interactive dialogue with LLM could be found in Appendix S.5.

The next material under examination in our case study is **paclitaxel** (PTX, $C_{47}H_{51}NO_{14}$), a compound of profound significance in the field of oncology and a critical component in the treatment of various cancers.



Now you are acting as an expert in chemistry and are celebrated for your exceptional proficiency in the intricate analysis of molecular formula, SMILES and molecular's function.

Example 1: Given the smiles: CC(C1=CC=CC=C1)N, what is the Molecular Formula? **Follow the chain-of-thought below**: 1) Start by decoding the Smiles notation to identify and count each type of atom. In the notation, "C" represents carbon, "N" represents nitrogen, and "H" (though not explicitly mentioned) represents hydrogen. 2) From the Smiles, we can count... 3) Now, assemble these counts into a molecular formula: C8H11N. **Example 2**... **Example 3**...

⋮

Example 1: Given the substrate: indole and acrylaldehyde, what is the product under the catalysis of MacMillan's imidazolidinone 2nd generation catalyst? **Follow the chain-of-thought below**: 1) MacMillan's imidazolidinone 2nd generation catalyst, can catalyze the 1,4-addition reaction of indole to α,β-unsaturated aldehydes. In this reaction... 2) So, for substrates indole and acrylaldehyde, the product in the catalytic reaction is 3-(1H-indol-3-yl)propanal. **Example 2**... **Example 3**...

User
Molecular Formula
Given the SMILES: CC(C)(C)C1NC(C(=O)N1C)CC2=CC=CC=C2, let's determine the **Molecular Formula**.

LLM
1) Decode the SMILES notation to identify and count each type of atom.
2) From the SMILES, we can count: 15 carbon atoms (C), 2 nitrogen atoms (N), 1 oxygen atom (O). The hydrogens are implicit in the SMILES notation, but based on standard valences and the structure, there are: 1 degrees of unsaturation from the =O, 4 degree of unsaturation from the aromatic ring (CC2=CC=CC=C2), 1 degree of unsaturation from the saturated ring (C1NC(C(=O)N1C)), and each "N" owns 1 less "H" than "C". So we can deduce there are (15+2)*2+2-2*6-2=22 hydrogens (H).
3) Assemble these counts into a molecular formula: C15H22N2O. ✓

User
Molecular Weight
Given the Molecular Formula: C15H22N2O, let's determine the **Molecular Weight**.

LLM
1) Use the atomic weights: Carbon (C) is approximately 12.01 g/mol, Hydrogen (H) is about 1.008 g/mol...
2) Calculate the contribution of each atom type: Carbon's is 15 x 12.01 = 180.15 g/mol, Hydrogen's is 22 x 1.008 = 22.176 g/mol...
3) Sum these values: 180.15 + 22.176 + 28.02 + 16.00 = 246.346 g/mol. ✓

User
H-bond Acceptor
Given the Molecular Formula: C15H22N2O and SMILES: CC(C)(C)C1NC(C(=O)N1C)CC2=CC=CC=C2, let's determine the **Number of H-bond Acceptors**.

LLM
1) We start by understanding what constitutes an H-bond acceptor. Hydrogen bond acceptors are typically atoms...
2) Now, let's analyze the given SMILES notation: CC(C)(C)C1NC(C(=O)N1C)CC2=CC=CC=C2. Looking at the SMILES, we can identify the potential H-bond acceptors: We see an oxygen atom in the carbonyl group (=O), which can act as an H-bond acceptor. Additionally, there are two nitrogen atoms (N) connected to carbons. Nitrogen atoms can also act as H-bond acceptors if they have lone pairs available...
3) Therefore, based on the SMILES notation, the molecule has 3 H-bond acceptors: one from the oxygen atom and two from the nitrogen atoms. ✓

User
Catalyst Product
Given the substrate: indole and (E)-hex-2-enal, **the product under the catalysis** of MacMillan's imidazolidinone 2nd generation catalyst is 3-(1H-indol-3-yl)hexanal.

LLM
1) MacMillan's imidazolidinone 2nd generation catalyst is known for its efficiency in catalyzing the 1,4-addition reaction of indole to α,β-unsaturated aldehydes.
2) In this type of reaction, the 4' carbon of the unsaturated aldehyde forms a new bond with the 3' carbon of the indole, which is the carbon on the five-membered ring of indole that is not adjacent to the nitrogen atom.The 3,4-double bond of the unsaturated aldehyde is broken, and new bonds are formed to create a new molecule.
3) When indole reacts with (E)-hex-2-enal, the product formed is 3-(1H-indol-3-yl)hexanal, where the indole moiety is attached to the hexanal chain at the 3-position of the indole ring. ✓

Fig. 11. Prompt engineering case study on MacMillan's imidazolidinone 2nd generation catalyst.

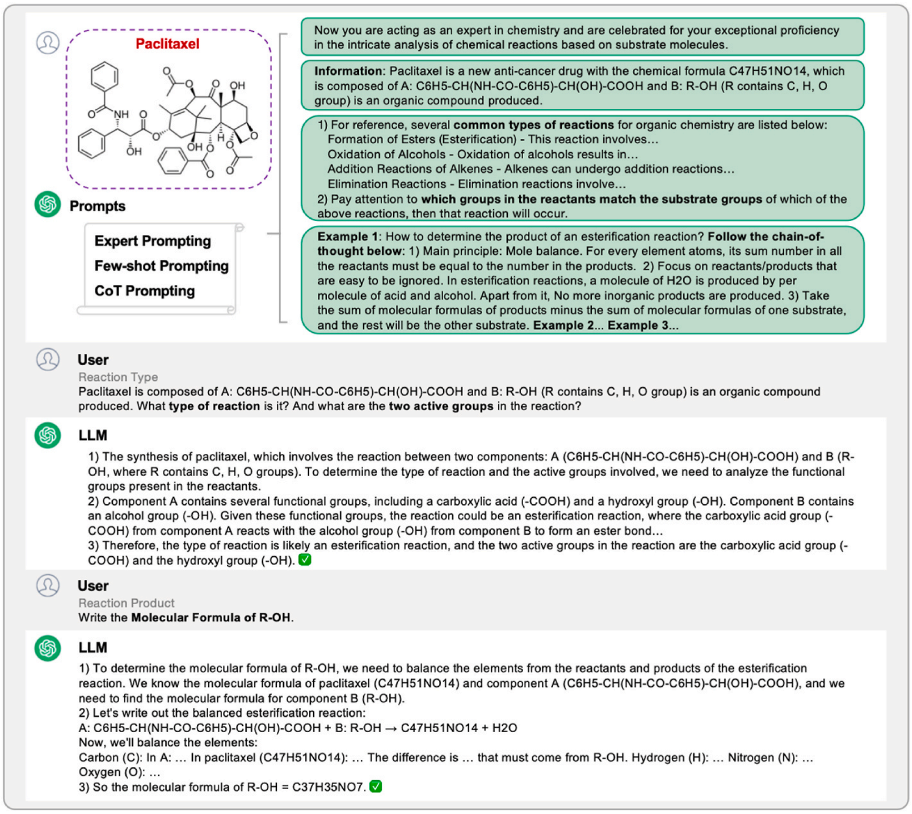
Paclitaxel's discovery and subsequent development mark a pivotal moment in the history of cancer therapy, as it introduced a novel mechanism of action that targets microtubules, thereby inhibiting the growth and division of cancer cells. Its efficacy in the treatment of breast, ovarian, and other cancers has established paclitaxel as a cornerstone in the chemotherapy arsenal [63]. The importance of paclitaxel extends beyond its direct clinical applications; it has also served as a template for the development of other taxane derivatives and has been a subject of extensive research in organic synthesis [64]. The complex structure of paclitaxel presents a significant challenge in the synthesis process, leading to the development of various strategies to improve yield, reduce cost, and enhance accessibility to this life-saving compound.

In this prompt engineering case study, we focus on a crucial step in the synthesis of an active intermediate of paclitaxel. Our objective is to assess the ability of LLMs to analyze and provide insights into the pathway of organic synthesis. By utilizing our domain-knowledge embedded prompt engineering method, we aim to demonstrate the potential of LLMs in assisting chemists in the design and optimization of synthetic routes for complex molecules, such as paclitaxel and its derivatives, thereby contributing to the advancement of both chemical research and pharmaceutical development.

As shown in Fig. 12, by utilizing custom-designed prompts, the LLM adeptly dissected a critical step in the synthesis of paclitaxel. It not only identified the reactive groups within the substrates that are capable of engaging in the chemical transformation but also correctly discerned the type of reaction and reconstructed the entire reaction scheme. This accomplishment underscores the LLM's potential in providing guidance for the synthesis of chemical compounds, suggesting that such models could play a pivotal role in streamlining the process of chemical synthesis, offering insights into reaction of complex molecules. This has implications for the advancement of medicinal chemistry and the development of pharmaceuticals, where efficient synthesis routes are of paramount importance. The complete interactive dialogue with LLM could be found in Appendix S.5.

In the concluding case study, we examine **lithium cobalt oxide** ($LiCoO_2$), a material of great importance in lithium-ion battery technology. Recognized by the 2018 Nobel Prize in Chemistry, $LiCoO_2$'s contribution to energy storage has been transformative, enabling the widespread use of portable electronics and electric vehicles [65]. As a cathode material, $LiCoO_2$ offers high energy density and stability, although research continues to address its lifecycle, cost, and environmental footprint.

In this prompt engineering case study, we delve into the analysis of $LiCoO_2$ crystals and their application advantages. We aim to harness the capabilities of LLMs to provide detailed insights into the crystallographic properties, electrochemical behavior, and optimization strategies for $LiCoO_2$. By employing our domain-knowledge embedded prompt engineering method, we expect to demonstrate the potential of LLMs in aiding researchers in the design and refinement of battery materials, thereby contributing to the progress of energy storage technologies and supporting the global transition towards sustainable



Now you are acting as an expert in chemistry and are celebrated for your exceptional proficiency in the intricate analysis of chemical reactions based on substrate molecules.

Information: Paclitaxel is a new anti-cancer drug with the chemical formula C47H51NO14, which is composed of A: C6H5-CH(NH-CO-C6H5)-CH(OH)-COOH and B: R-OH (R contains C, H, O group) is an organic compound produced.

1) For reference, several **common types of reactions** for organic chemistry are listed below:
Formation of Esters (Esterification) - This reaction involves…
Oxidation of Alcohols - Oxidation of alcohols results in…
Addition Reactions of Alkenes - Alkenes can undergo addition reactions…
Elimination Reactions - Elimination reactions involve…
2) Pay attention to **which groups in the reactants match the substrate groups** of which of the above reactions, then that reaction will occur.

Example 1: How to determine the product of an esterification reaction? **Follow the chain-of-thought below**: 1) Main principle: Mole balance. For every element atoms, its sum number in all the reactants must be equal to the number in the products. 2) Focus on reactants/products that are easy to be ignored. In esterification reactions, a molecule of H2O is produced by per molecule of acid and alcohol. Apart from it, No more inorganic products are produced. 3) Take the sum of molecular formulas of products minus the sum of molecular formulas of one substrate, and the rest will be the other substrate. **Example 2**… **Example 3**…

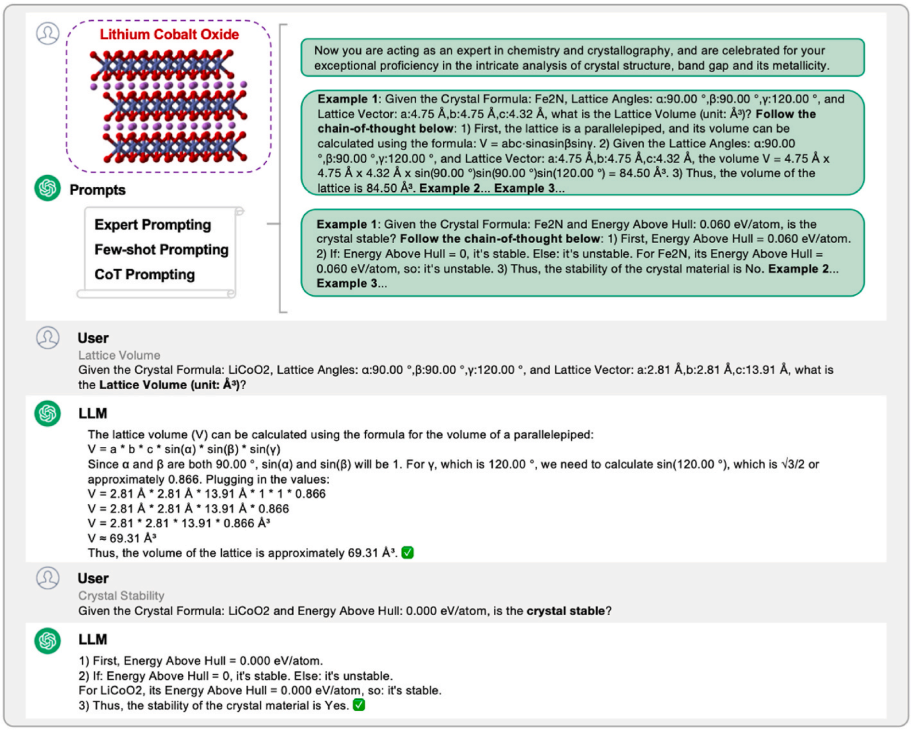
User
Reaction Type
Paclitaxel is composed of A: C6H5-CH(NH-CO-C6H5)-CH(OH)-COOH and B: R-OH (R contains C, H, O group) is an organic compound produced. What **type of reaction** is it? And what are the **two active groups** in the reaction?

LLM
1) The synthesis of paclitaxel, which involves the reaction between two components: A (C6H5-CH(NH-CO-C6H5)-CH(OH)-COOH) and B (R-OH, where R contains C, H, O groups). To determine the type of reaction and the active groups involved, we need to analyze the functional groups present in the reactants.
2) Component A contains several functional groups, including a carboxylic acid (-COOH) and a hydroxyl group (-OH). Component B contains an alcohol group (-OH). Given these functional groups, the reaction could be an esterification reaction, where the carboxylic acid group (-COOH) from component A reacts with the alcohol group (-OH) from component B to form an ester bond…
3) Therefore, the type of reaction is likely an esterification reaction, and the two active groups in the reaction are the carboxylic acid group (-COOH) and the hydroxyl group (-OH). ✅

User
Reaction Product
Write the **Molecular Formula of R-OH**.

LLM
1) To determine the molecular formula of R-OH, we need to balance the elements from the reactants and products of the esterification reaction. We know the molecular formula of paclitaxel (C47H51NO14) and component A (C6H5-CH(NH-CO-C6H5)-CH(OH)-COOH), and we need to find the molecular formula for component B (R-OH).
2) Let's write out the balanced esterification reaction:
A: C6H5-CH(NH-CO-C6H5)-CH(OH)-COOH + B: R-OH → C47H51NO14 + H2O
Now, we'll balance the elements:
Carbon (C): In A: … In paclitaxel (C47H51NO14): … The difference is … that must come from R-OH. Hydrogen (H): … Nitrogen (N): … Oxygen (O): …
3) So the molecular formula of R-OH = C37H35NO7. ✅

Fig. 12. Prompt engineering case study on paclitaxel.



Now you are acting as an expert in chemistry and crystallography, and are celebrated for your exceptional proficiency in the intricate analysis of crystal structure, band gap and its metallicity.

Example 1: Given the Crystal Formula: Fe2N, Lattice Angles: α:90.00 °,β:90.00 °,γ:120.00 °, and Lattice Vector: a:4.75 Å,b:4.75 Å,c:4.32 Å, what is the Lattice Volume (unit: Å³)? **Follow the chain-of-thought below**: 1) First, the lattice is a parallelepiped, and its volume can be calculated using the formula: V = abc·sinαsinβsinγ. 2) Given the Lattice Angles: α:90.00 °,β:90.00 °,γ:120.00 °, and Lattice Vector: a:4.75 Å,b:4.75 Å,c:4.32 Å, the volume V = 4.75 Å x 4.75 Å x 4.32 Å x sin(90.00 °)sin(90.00 °)sin(120.00 °) = 84.50 Å³. 3) Thus, the volume of the lattice is 84.50 Å³. **Example 2**… **Example 3**…

Example 1: Given the Crystal Formula: Fe2N and Energy Above Hull: 0.060 eV/atom, is the crystal stable? **Follow the chain-of-thought below**: 1) First, Energy Above Hull = 0.060 eV/atom. 2) If: Energy Above Hull = 0, it's stable. Else: it's unstable. For Fe2N, its Energy Above Hull = 0.060 eV/atom, so: it's unstable. 3) Thus, the stability of the crystal material is No. **Example 2**… **Example 3**…

User
Lattice Volume
Given the Crystal Formula: LiCoO2, Lattice Angles: α:90.00 °,β:90.00 °,γ:120.00 °, and Lattice Vector: a:2.81 Å,b:2.81 Å,c:13.91 Å, what is the **Lattice Volume (unit: Å³)**?

LLM
The lattice volume (V) can be calculated using the formula for the volume of a parallelepiped:
V = a * b * c * sin(α) * sin(β) * sin(γ)
Since α and β are both 90.00 °, sin(α) and sin(β) will be 1. For γ, which is 120.00 °, we need to calculate sin(120.00 °), which is √3/2 or approximately 0.866. Plugging in the values:
V = 2.81 Å * 2.81 Å * 13.91 Å * 1 * 1 * 0.866
V = 2.81 Å * 2.81 Å * 13.91 Å * 0.866
V = 2.81 * 2.81 * 13.91 * 0.866 Å³
V ≈ 69.31 Å³
Thus, the volume of the lattice is approximately 69.31 Å³. ✅

User
Crystal Stability
Given the Crystal Formula: LiCoO2 and Energy Above Hull: 0.000 eV/atom, is the **crystal stable**?

LLM
1) First, Energy Above Hull = 0.000 eV/atom.
2) If: Energy Above Hull = 0, it's stable. Else: it's unstable.
For LiCoO2, its Energy Above Hull = 0.000 eV/atom, so: it's stable.
3) Thus, the stability of the crystal material is Yes. ✅

Fig. 13. Prompt engineering case study on lithium cobalt oxide.

energy solutions.

In this case study, the LLM meticulously analyzed the fundamental properties of lithium cobalt oxide ($LiCoO_2$), accurately determining its lattice volume and stability, as shown in Fig. 13. This achievement is of significant importance in the field of crystallography and future development of lithium-ion battery technologies. The complete interactive dialogue with LLM could be found in Appendix S.5.

4. Conclusion and future directions

The integration of domain-specific knowledge into prompt engineering has demonstrated its effectiveness in enhancing the performance of LLMs across various tasks in chemistry, materials science, and biology. Our proposed domain-knowledge embedded prompt engineering method outperforms traditional generic prompt engineering strategies on metrics such as capability, accuracy, F1 score, and hallucination drop. The incorporation of domain expertise into prompts not only guides the LLM to synthesize more relevant knowledge but also provides a clear reasoning path for complex tasks. Our case studies further validate the effectiveness of this approach in analyzing intricate materials like the MacMillan catalyst, paclitaxel, and $LiCoO_2$, demonstrating the potential of LLMs to assist experts in molecular design and optimization when equipped with domain-specific prompts. The complete code implementation of our work is listed in Appendix S.1.

Limitations and potential future directions of our work is also concluded below:

Expansion of Domain Coverage: While our study has focused on chemistry, materials, and biology, the concept of domain-knowledge embedded prompt engineering can be extended to other scientific domains. Future work can explore the development of tailored prompts for fields such as physics, geology, and medicine to unlock the full potential of LLMs in diverse scientific applications.

Integration of Datasets and Tools: To further enhance the reasoning capabilities of LLMs, future prompt engineering can integrate external datasets and domain-specific tools. Linking prompts to chemical databases, computational chemistry software, or biological sequence analysis tools could enable the LLM to leverage additional information for more accurate predictions.

Multi-Modal Prompting: Incorporating visual information, such as molecular structures or crystal images, into prompts can provide a more intuitive understanding for LLMs. Multi-modal prompting techniques combining textual and visual cues can potentially lead to even stronger performance gains.

Human-in-the-Loop Refinement: Iteratively refining prompts with input from domain experts can help to uncover more effective prompting strategies. Human-in-the-loop systems that leverage the complementary strengths of LLMs and human experts have the potential to achieve highly optimized prompts.

Prompt Engineering Benchmarking: To ensure comprehensive and fair evaluation of prompting strategies, it is meaningful to establish standardized benchmarks across multiple LLMs, especially the recently released ones. This approach allows researchers to compare the performance of prompt engineering on different LLMs, thereby driving innovation in the field. Creating diverse datasets with a wide range of tasks and molecules will enable robust evaluation and facilitate the development of more effective prompting techniques for various LLMs.

In summary, domain-knowledge embedded prompt engineering has shown great promise for unlocking the potential of LLMs in scientific domains. By integrating domain expertise into prompts, LLMs can generate more accurate and contextually relevant responses. As prompt engineering techniques continue to evolve, LLMs have the potential to become powerful allies for scientists, assisting in the exploration and discovery of new materials, molecules, and biological entities.

CRediT authorship contribution statement

Hongxuan Liu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Haoyu Yin:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing – original draft, Writing – review & editing. **Zhiyao Luo:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xiaonan Wang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgement

This work is supported by the National Key R&D Program of China (No. 2022ZD0117501).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2024.07.004.

References

[1] Wang S, Fu J, Liu Y, Saravanan RS, Luo J, Deng S, Sham T-K, Sun X, Mo Y. Design principles for sodium superionic conductors. Nat Commun 2023;14(1):7615.
[2] Dubey DK, Thakur D, Yadav RAK, Ram Nagar M, Liang T-W, Ghosh S, Jou J-H. High-throughput virtual screening of host materials and rational device engineering for highly efficient solution-processed organic light-emitting diodes. ACS Appl Mater Interfaces 2021;13(22):26204–17.
[3] Medasani B, Gamst A, Ding H, Chen W, Persson KA, Asta M, Canning A, Haranczyk M. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. npj Comput Mater 2016;2(1):1.
[4] Fetanat M, Keshtiara M, Keyikoglu R, Khataee A, Daiyan R, Razmjou A. Machine learning for design of thin-film nanocomposite membranes. Separ Purif Technol 2021;270:118383.
[5] Goebel R, Skiborowski M. Machine-based learning of predictive models in organic solvent nanofiltration: pure and mixed solvent flux. Separ Purif Technol 2020;237: 116363.
[6] Guan J, Huang T, Liu W, Feng F, Japip S, Li J, Wu J, Wang X, Zhang S. Design and prediction of metal organic framework-based mixed matrix membranes for CO2 capture via machine learning. Cell Reports Physical Science 2022;3(5):100864.
[7] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. Nature 2020;577(7792):706–10.
[8] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[9] Yang Y. A machine-learning prediction method of lithium-ion battery life based on charge process for different applications. Appl Energy 2021;292:116897.
[10] Fei Z, Yang F, Tsui K-L, Li L, Zhang Z. Early prediction of battery lifetime via a machine learning based framework. Energy 2021;225:120205.
[11] Schütt KT, Gastegger M, Tkatchenko A, Müller K-R, Maurer RJ. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. Nat Commun 2019;10(1):5024.
[12] Stöhr M, Medrano Sandonas L, Tkatchenko A. Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks. J Phys Chem Lett 2020;11(16):6835–43.
[13] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel CJ, Ceder G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nat Mach Intell 2023;5(9):1031–41.
[14] Li H, Tang Z, Gong X, Zou N, Duan W, Xu Y. Deep-learning electronic-structure calculation of magnetic superstructures. Nature Computational Science 2023;3(4): 321–7.
[15] Ziatdinov M, Dyck O, Maksov A, Li X, Sang X, Xiao K, Unocic RR, Vasudevan R, Jesse S, Kalinin Sv. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. ACS Nano 2017;11(12):12742–52.
[16] Li J, Telychko M, Yin J, Zhu Y, Li G, Song S, Yang H, Li J, Wu J, Lu J, Wang X. Machine vision automated chiral molecule detection and classification in molecular imaging. J Am Chem Soc 2021;143(27):10177–88.
[17] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. Science 2018;361(6400): 360–5.
[18] Gebauer NWA, Gastegger M, Hessmann SSP, Müller K-R, Schütt KT. Inverse design of 3d molecular structures with conditional generative neural networks. Nat Commun 2022;13(1):973.
[19] Weiss T, Mayo Yanes E, Chakraborty S, Cosmo L, Bronstein AM, Gershoni-Poranne R. Guided diffusion for inverse molecular design. Nature Computational Science 2023;3(10):873–82.

[20] Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, Li A, Cubillos-Ruiz A, Krishnan A, Jin W, Manson AL, Friedrichs J, Helbig R, Hajian B, Fiejtek DK, Wagner FF, Soutter HH, Earl AM, Stokes JM, Collins JJ. Discovery of a structural class of antibiotics with explainable deep learning. Nature 2024;626(7997): 177–85.

[21] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Baker D. De novo design of protein structure and function with RFdiffusion. Nature 2023;620(7976): 1089–100.

[22] Vázquez Torres S, Leung PJY, Venkatesh P, Lutz ID, Hink F, Huynh H-H, Becker J, Yeh AH-W, Juergens D, Bennett NR, Hoofnagle AN, Huang E, MacCoss MJ, Expòsit M, Lee GR, Bera AK, Kang A, de La Cruz J, Levine PM, Baker D. De novo design of high-affinity binders of bioactive helical peptides. Nature 2024;626 (7998):435–42.

[23] Burger B, Maffettone PM, Gusev V v, Aitchison CM, Bai Y, Wang X, Li X, Alston BM, Li B, Clowes R, Rankin N, Harris B, Sprick RS, Cooper AI. A mobile robotic chemist. Nature 2020;583(7815):237–41.

[24] Nega PW, Li Z, Ghosh V, Thapa J, Sun S, Hartono NTP, Nellikkal MAN, Norquist AJ, Buonassisi T, Chan EM, Schrier J. Using automated serendipity to discover how trace water promotes and inhibits lead halide perovskite crystal formation. Appl Phys Lett 2021;119(4):041903.

[25] Bannwarth C, Ehlert S, Grimme S. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. J Chem Theor Comput 2019;15(3):1652–71.

[26] Jha G, Heine T. Spin–orbit coupling corrections for the GFN-xTB method. J Chem Phys 2023;158(4):044120.

[27] Zeng J, Zhang D, Lu D, Mo P, Li Z, Chen Y, Rynik M, Huang L, Li Z, Shi S, Wang Y, Ye H, Tuo P, Yang J, Ding Y, Li Y, Tisi D, Zeng Q, Bao H, Wang H. DeePMD-kit v2: a software package for deep potential models. J Chem Phys 2023;159(5):54801.

[28] Slattery A, Wen Z, Tenblad P, Sanjosé-Orduna J, Pintossi D, den Hartog T, Noël T. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. Science 2024;383(6681):eadj1817.

[29] Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, McDermott MJ, Gallant M, Cubuk ED, Merchant A, Kim H, Jain A, Bartel CJ, Persson K, Zeng Y, Ceder G. An autonomous laboratory for the accelerated synthesis of novel materials. Nature 2023;624(7990):86–91.

[30] Xu S, Li J, Cai P, Liu X, Liu B, Wang X. Self-improving photosensitizer discovery system via bayesian search with first-principle simulations. J Am Chem Soc 2021; 143(47):19769–77.

[31] Gao H, Zhong S, Zhang W, Igou T, Berger E, Reid E, Zhao Y, Lambeth D, Gan L, Afolabi MA, Tong Z, Lan G, Chen Y. Revolutionizing membrane design using machine learning-bayesian optimization. Environ Sci Technol 2022;56(4): 2572–81.

[32] Rao Z, Tung P-Y, Xie R, Wei Y, Zhang H, Ferrari A, Klaver TPC, Körmann F, Sukumar PT, Kwiatkowski da Silva A, Chen Y, Li Z, Ponge D, Neugebauer J, Gutfleisch O, Bauer S, Raabe D. Machine learning–enabled high-entropy alloy discovery. Science 2022;378(6615):78–85.

[33] Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. ChemCrow: augmenting large-language models with chemistry tools. 2023. ArXiv E-Prints, arXiv:2304.05376.

[34] Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 2022;4(12):1256–64.

[35] Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature 2023;624(7990):80–5.

[36] Wei J, Wang X, Schuurmans D, Bosma M, ichter brian, Xia F, Chi E, Le Q v, Zhou D. Chain-of-Thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in neural information processing systems, vol. 35. Curran Associates, Inc; 2022. p. 24824–37.

[37] Ahmed T, Devanbu P. Few-shot training LLMs for project-specific code-summarization. Proceedings of the 37th IEEE/ACM international conference on automated software engineering. 2023.

[38] White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC. A prompt pattern catalog to enhance prompt engineering with ChatGPT. 2023. ArXiv E-Prints, arXiv:2302.11382.

[39] Zhou Y, Ioan Muresanu A, Han Z, Paster K, Pitis S, Chan H, Ba J. Large Language models are human-level prompt engineers. 2022. ArXiv E-Prints, arXiv: 2211.01910.

[40] Ekin S. Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. https://doi.org/10.36227/techrxiv.22683919.v2; 2023.

[41] Xie T, Wan Y, Huang W, Yin Z, Liu Y, Wang S, Linghu Q, Kit C, Grazian C, Zhang W, Razzak I, Hoex B. Darwin series: domain specific large language models for natural science. 2023. ArXiv E-Prints, arXiv:2308.13565.

[42] Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023;51(12):2629–33.

[43] Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638.

[44] Schreiber SL. Organic synthesis toward small-molecule probes and drugs. Proc Natl Acad Sci USA 2011;108(17):6699–702.

[45] Kirk O, Borchert TV, Fuglsang CC. Industrial enzyme applications. Curr Opin Biotechnol 2002;13(4):345–51.

[46] Sharma A, Gupta G, Ahmad T, Mansoor S, Kaur B. Enzyme engineering: current trends and future perspectives. Food Rev Int 2021;37(2):121–54.

[47] Surek T. Crystal growth and materials research in photovoltaics: progress and challenges. J Cryst Growth 2005;275(1):292–304.

[48] Zhang C, Zhang J, Ma X, Feng Q. Semiconductor photovoltaic cells. Springer; 2021.

[49] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. 2020. ArXiv E-Prints, arXiv: 2009.03300.

[50] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, Santoro A, Gupta A, Garriga-Alonso A, Kluska A, Lewkowycz A, Agarwal A, Power A, Ray A, Warstadt A, Kocurek AW, Safaya A, Tazarv A, Wu Z. Beyond the Imitation Game: quantifying and extrapolating the capabilities of language models. ArXiv E-Prints, arXiv:2206; 2022, 04615.

[51] Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J. Training verifiers to solve math word problems. ArXiv E-prints. 2021. arXiv:2110.14168.

[52] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem substance and compound databases. Nucleic Acids Res 2016;44(D1):D1202–13.

[53] Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.

[54] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. Apl Mater 2013;1(1): 011002.

[55] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Amodei D. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems, vol. 33. Curran Associates, Inc.; 2020. p. 1877–901.

[56] Zhang SJ, Florin S, Lee AN, Niknafs E, Marginean A, Wang A, Tyser K, Chin Z, Hicke Y, Singh N, Udell M, Kim Y, Buonassisi T, Solar-Lezama A, Drori I. Exploring the MIT mathematics and EECS curriculum using large language models. 2023. ArXiv E-Prints, arXiv:2306.08997.

[57] Chu Z, Chen J, Chen Q, Yu W, He T, Wang H, Peng W, Liu M, Qin B, Liu T. A survey of chain of thought reasoning: advances, frontiers and future. ArXiv E-Prints; 2023. arXiv:2309.15402.

[58] OpenAI. OpenAI's documentation for language models. https://platform.openai.com/docs/models/models; 2023.

[59] Imani S, Du L, Shrivastava H. Mathprompter: mathematical reasoning using large language models. 2023. *arXiv preprint* arXiv:2303.05398.

[60] Wu Y, Jia F, Zhang S, Li H, Zhu E, Wang Y, Wang C. MathChat: converse to tackle challenging math problems with LLM agents. In: In ICLR 2024 workshop on large language model (LLM) agents; 2024.

[61] MacMillan DWC. The advent and development of organocatalysis. Nature 2008; 455(7211):304–8.

[62] Deepa, Singh S. Recent development of recoverable MacMillan catalyst in asymmetric organic transformations. Adv Synth Catal 2021;363(3):629–56.

[63] Markman M, Mekhail TM. Paclitaxel in cancer therapy. Expet Opin Pharmacother 2002;3(6):755–66.

[64] Mosca L, Ilari A, Fazi F, Assaraf YG, Colotti G. Taxanes in cancer treatment: activity, chemoresistance and its overcoming. Drug Resist Updates 2021;54: 100742.

[65] Wu Q, Zhang B, Lu Y. Progress and perspective of high-voltage lithium cobalt oxide in lithium-ion batteries. J Energy Chem 2022;74:283–308.