

Perovskite-R1: A Domain-Specialized LLM for Intelligent Discovery of Precursor Additives and Experimental Design

Xin-De Wang^{1,†}, Zhi-Rui Chen^{2,†}, Peng-Jie Guo¹, Ze-Feng Gao^{1,*}, Cheng Mu^{2,*}, Zhong-Yi Lu^{1,*}

¹School of Physics, Renmin University of China, Beijing, China

²School of Chemistry and Life Resource, Renmin University of China, Beijing, China

[†]Equally contributed

^{*}Corresponding authors

Abstract

Perovskite solar cells (PSCs) have rapidly emerged as a leading contender in next-generation photovoltaic technologies, owing to their exceptional power conversion efficiencies and advantageous material properties. Despite these advances, challenges such as long-term stability, environmental sustainability, and scalable manufacturing continue to hinder their commercialization. Precursor additive engineering has shown promise in addressing these issues by enhancing both the performance and durability of PSCs. However, the explosive growth of scientific literature and the complex interplay of materials, processes, and device architectures make it increasingly difficult for researchers to efficiently access, organize, and utilize domain knowledge in this rapidly evolving field. To address this gap, we introduce Perovskite-R1, a specialized large language model (LLM) with advanced reasoning capabilities tailored for the discovery and design of PSC precursor additives. By systematically mining and curating 1,232 high-quality scientific publications and integrating a comprehensive library of 33,269 candidate materials, we constructed a domain-specific instruction-tuning dataset using automated question-answer generation and chain-of-thought reasoning. Fine-tuning the QwQ-32B model on this dataset resulted in Perovskite-R1, which can intelligently synthesize literature insights and generate innovative and practical solutions for defect passivation and the selection of precursor additives. Experimental validation of several model-proposed strategies confirms their effectiveness in improving material stability and performance. Our work demonstrates the potential of domain-adapted LLMs in accelerating materials discovery and provides a closed-loop framework for intelligent, data-driven advancements in perovskite photovoltaic research.

Keywords: large language model, perovskite

Introduction

In recent years, perovskite solar cells (PSCs) have garnered significant attention from both academia and industry due to their tremendous potential in sustainable energy applications. Since their introduction in 2009, the power conversion efficiency (PCE) of PSCs has rapidly increased from 3.8% to 26.95% [1, 2], highlighting their promising prospects as next-generation photovoltaic

technology. The remarkable efficiency improvement of PSCs is mainly attributed to the unique advantages of perovskite materials, such as highly ordered crystal structures, excellent light absorption, tunable band gaps, and cost-effective solution-based fabrication [3, 4]. These features not only enable high performance but also facilitate the scalability of PSCs for large-scale applications.

However, the commercialization of PSCs still faces significant challenges, including long-term operational stability, environmental sustainability, and scalable manufacturing [5, 6]. Issues such as sensitivity to moisture and heat, material degradation, and process compatibility remain key bottlenecks limiting practical deployment. To address these challenges, precursor additive engineering has emerged as an effective strategy to enhance both the stability and efficiency of PSCs. By introducing suitable additives into the precursor solution, researchers can optimize crystallization, passivate defects, and improve film quality, thereby significantly improving device performance and durability. This approach offers strong support for the further industrialization of PSC technology.

Traditionally, progress in PSCs research has relied on researchers conducting comprehensive literature reviews and then applying intuition and experience-based judgment to analyze findings, followed by experimental validation. However, the rapid advancement of PSCs in recent years has led to an exponential increase in the number of related publications, making it increasingly challenging for researchers to efficiently access and utilize the ever-growing body of knowledge in this field. This challenge is particularly acute given the complex interplay among material composition, fabrication processes, and device architecture that characterize PSCs research. On the other hand, existing artificial intelligence systems [7–9] in materials science typically focus on specific prediction tasks or general scientific knowledge [10, 11], lacking the specialized capability to address the unique characteristics of PSC research [12]. This gap highlights the urgent need for an integrated system that can systematically organize domain knowledge and provide intelligent assistance to researchers.

Recently, large language models (LLMs) have developed rapidly, with numerous high-performance models emerging [13–17]. Advances in LLMs are primarily driven by increases in model size, the scale of pre-training data, and enhanced reasoning capabilities. Recently, much attention has focused on improving reasoning abilities, which not only boost accuracy but also enhance interpretability and controllability. For example, the introduction of Chain-of-Thought (CoT) prompting in 2022 led to significant improvements in mathematical reasoning tasks for large models [18], laying the groundwork for further developments in the field. Moreover, integrating CoT methods with reinforcement learning has further advanced model autonomy, enabling superior performance in tasks such as mathematics, programming, and scientific reasoning [19–21]. At the same time, applying LLMs to specialized domains has become an important trend. However, due to complex terminologies and extensive knowledge structures in these fields, general-purpose models often fall short. Techniques such as domain-specific fine-tuning and structured knowledge injection have facilitated the emergence of specialized LLMs. For instance, dedicated LLMs have been developed and applied in areas such as biomedicine, materials science, education, and finance [22–33].

In this work, we present Perovskite-R1, an LLM with advanced reasoning capabilities designed specifically for the discovery and design of PSC precursor additives. Perovskite-R1 aims to provide intelligent and systematic support for material design and screening in the field of perovskite photovoltaics. Leveraging the capabilities of Perovskite-R1, we systematically identify a set of potential defect compensation strategies for perovskite precursor additives. Several of these strategies are further validated through experimental investigations, which demonstrated that the solutions proposed by Perovskite-R1 are both rational and reliable, leading to improvements in material performance and stability. Specifically, we collect 1,232 scientific publications related to perovskite precursor design as well as a candidate library containing 33,269 materials. Utilizing the OpenAI o1

model¹[19], we transform the content of these papers into an instruction-tuning dataset in the form of question–answer pairs, thereby enriching the model’s domain-specific knowledge and practical examples. Building on this high-quality instruction dataset, we fine-tune the QwQ-32B pre-trained model²[34] to develop Perovskite-R1. Thanks to this comprehensive dataset and instruction tuning, Perovskite-R1 can effectively integrate existing knowledge on perovskite precursors with the latest research advances and experimental data, generating innovative and practical design solutions.

To experimentally validate the predictive power of our model, we conducted a systematic comparison between additives recommended by Perovskite-R1 (3, 5-difluoropyridine-2-carboxylic acid (AI-DFCA) and 5-hydroxy-2-methylbenzoic acid (AI-HMBA)) and those chosen manually by researchers (gallic acid (Manual-GA) and caffeic acid (Manual-CA)). All additives were incorporated at equal concentrations into $\text{Cs}_{0.05}\text{MA}_{0.1}\text{FA}_{0.85}\text{PbI}_3$ perovskite devices under identical fabrication conditions. The results revealed that model-identified additives significantly improved device performance, while manually selected additives led to inferior outcomes. This highlights the advantage of data-driven screening over traditional, experience-based approaches in complex materials discovery, opening new avenues for intelligent material discovery in the PSC field.

Related work

AI technology for materials science Research in materials science is characterized by a high degree of diversity and complexity. Consequently, the development of computational tools for material properties and the discovery of new materials have been a primary focus of research efforts. In light of the recent advancements in AI, it becomes imperative to explore the potential of AI technology in the domain of materials science [35–37].

A notable example of work on predicting material properties using AI is ALIGNN [38]. By incorporating bond angle information into graph neural networks, ALIGNN significantly improves the accuracy of material property predictions. Due to its excellent performance and ease of training, ALIGNN is still used in many studies today [39–42]. In addition to ALIGNN, many other efforts have been made to predict material properties [43–45]. Each method has its own strengths and weaknesses, and specialises in a particular area. To this day, material prediction methods applied to different fields are still being developed.

Another important application of AI in materials science is reverse design. This area has seen a lot of excellent work. An earlier work is Crystal Diffusion Variational Autoencoder (CDVAE), which uses a diffusion process to generate crystal structures with SE(3) and periodic invariance [46]. This process optimizes specified properties, such as final energy states. Another excellent work is DiffCSP [47]. It provides greater stability and accuracy in structural prediction by combining diffusion lattices and fractional coordinates, as well as introducing improvements to degeneracy, such as E(3) equivariance. GNoME is a significant work that combines graph neural networks (GNNs) and density functional theory to successfully predict more than 380,000 novel stabilized materials, thereby significantly expanding the number of known stabilized materials [48].

LLM for materials science In consideration of the recent advancements in LLM, there have been several attempts to apply this technique to the materials domain. LLMs exhibit greater flexibility and maneuverability in comparison to diffusion models and GNNs, suggesting significant potential in the field of materials research [35, 37].

¹The full version of OpenAI o1 model was released on December 5, 2024.

²The QwQ-32B model was first released on March 6, 2025.

A conventional approach involves training a materials-specific LLM from the scratch with the assistance of a materials database, as exemplified by CrystaLLM [25]. CrystaLLM is an autoregressive LLM that has been trained on millions of crystal structure (CIF) files to generate plausible inorganic crystal structures from textual prompts. Moreover, there are other studies that adopt a similar approach. These studies borrow fine-tuning techniques [26] or combine them with flow-matching models [27]. These techniques reduce the computational power requirement during training, enhance the model, and improve its capabilities.

In addition to the aforementioned works, numerous recent studies have adopted techniques and methodologies from the LLM field, thereby facilitating novel advancements in materials science. MatterChat employs a multimodal LLM that integrates the atomic structure of a material with the text, thereby facilitating property prediction and experimental process reasoning [49]. HoneyComb is a materials science-specific LLM intelligence system that integrates tool invocation, structured knowledge base, and retrieval modules to improve the accuracy and adaptability of complex tasks, such as computational tasks or documentation [50]. LLaMP is an integrated platform that combines a material project database and high-throughput simulation [51]. It supports crystal structure editing, molecular dynamics simulation calling, and high-precision property prediction. LLaMP is based on the Retrieval-augmented Generation (RAG) and ReAct agents, which significantly reduces the phantom problem.

Results

We train a domain-specific LLM, Perovskite-R1, focusing on defect passivation in the field of perovskite materials, with the goal of enhancing its scientific reasoning and generation capabilities for tasks such as the selection of precursor additives and experimental design. To achieve this, we develop a comprehensive framework encompassing high-quality dataset construction, supervised fine-tuning with CoT annotations, structured prompt engineering, and systematic performance evaluation. As illustrated in Figure 1, the framework integrates curated perovskite literature and a drug-like compound library, ensuring both domain coverage and improved interpretability and generalization in complex materials science tasks. We further construct a benchmark dataset and compare Perovskite-R1 against several leading LLMs, demonstrating its superior performance in addressing perovskite-related challenges. Finally, we showcase the application of Perovskite-R1 in discovering high-performance defect-passivating precursor additives, highlighting its potential as an intelligent engine for next-generation materials discovery. The following subsections provide a detailed account of these key components and results.

Framework of Perovskite-R1

The overall construction process of Perovskite-R1 can be divided into four key stages. First, a high-quality instruction fine-tuning dataset is built through systematic literature mining. Second, based on this dataset, the LLM is fine-tuned using the LoRA strategy, resulting in the Perovskite-R1 model with specialized expertise in perovskite materials. Third, structured prompts are designed to guide the model in specific tasks such as experimental design and material screening. Finally, the model-generated recommendations are implemented as practical experimental protocols, which are then synthesized and validated in the laboratory, thus establishing a closed-loop system from language model generation to physical experimentation.

The foundation of Perovskite-R1 lies in the construction of a high-quality instruction fine-tuning dataset specifically oriented toward the task of defect passivation in perovskite precursors. To achieve this, we meticulously curate a selection of 1,232 peer-reviewed research articles from top

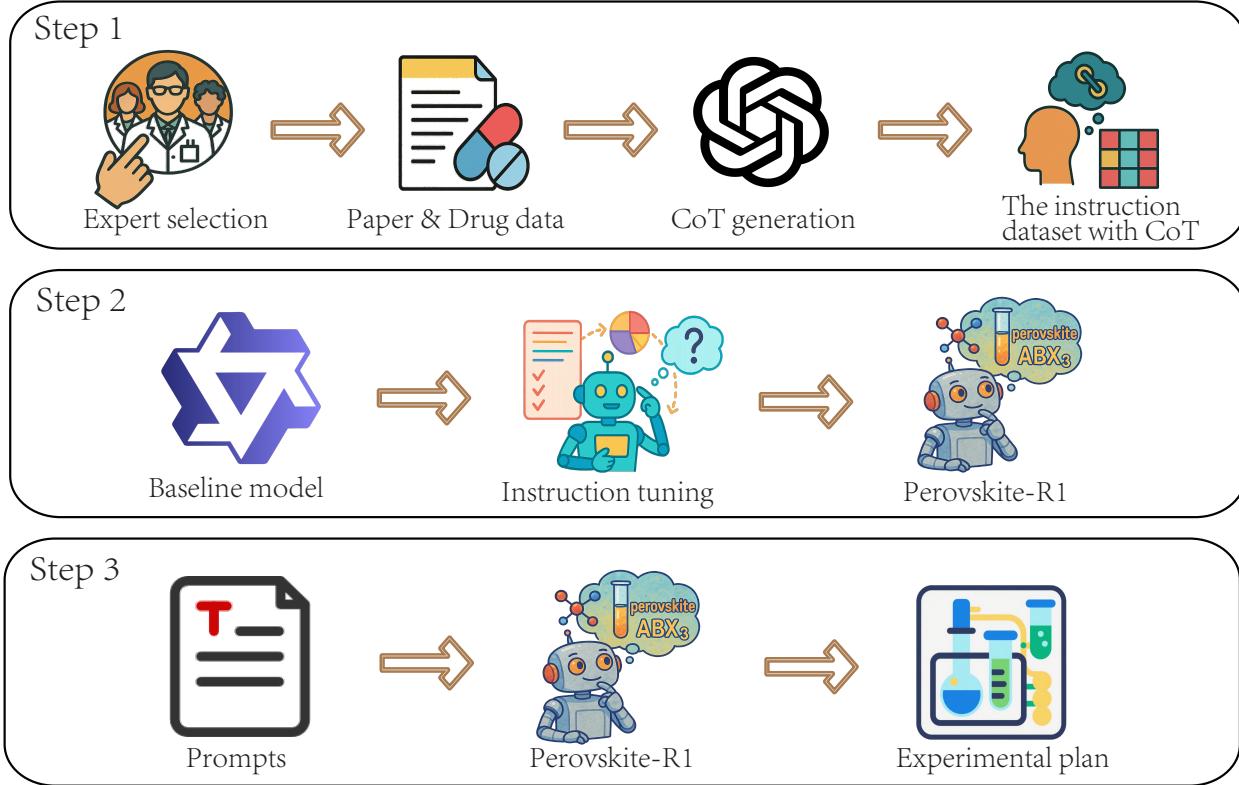


Figure 1: The detail of the construction of the proposed Perovskite-R1. In step 1, we select the proper papers and drug data, and then use an LLM to construct a dataset containing CoT. In step 2, we use scientific journal articles and drug data to conduct instruction tuning the baseline model. Finally, in step 3, we design prompts to obtain the desired experimental protocols.

scientific publishers, including the American Chemical Society (ACS), Elsevier, the Royal Society of Chemistry (RSC), Wiley, and Springer Nature. Our selection criteria prioritize studies that provided detailed descriptions of precursor additive synthesis pathways, material characterization methods, and performance optimization strategies, with a particular emphasis on works demonstrating significant achievements in device efficiency or stability and publications featured in high-impact journals. To ensure comprehensive coverage of both foundational and cutting-edge knowledge, we trace the evolution of perovskite precursor additive strategies from their early development to the latest breakthroughs, thus capturing the full temporal spectrum of the field. Additionally, to enhance the model’s generalization across chemical space, we incorporate a large-scale drug-like molecular library containing 33,269 distinct compounds. This resource not only furnishes the model with a diverse array of candidate additive structures, but also enables it to learn the intricate relationships between molecular structures and material properties, thereby strengthening its chemical reasoning and generative capabilities.

During the data preprocessing stage, we utilize the pdfminer tool to convert the selected PDF documents into structured plain text. To balance contextual coherence and computational efficiency, each paper is segmented into text chunks of up to 2,500 tokens, with a 20% overlap between adjacent segments to maintain semantic consistency across paragraphs. This strategy not only enhances the logical integrity of the text sequences but also significantly improves the coherence and quality of CoT reasoning paths in subsequent question–answer generation tasks. Next, we employ the OpenAI o1 API to automatically generate high-quality question–answer pairs from these text segments.

To further strengthen the model’s scientific reasoning abilities, we introduce a customized CoT template that guides the model to produce responses containing explicit reasoning steps along with final conclusions. As a result, the original corpus is transformed into an instruction-tuning dataset with deep semantic structure and scientific logic, endowing the model with advanced capabilities in experimental design and mechanistic understanding of materials.

For the model training stage, we select the powerful QwQ-32B model [34] as the base language model and perform efficient parameter fine-tuning using the LoRA technique on the aforementioned CoT-enhanced dataset. By introducing low-rank matrices for targeted adjustments to the original model parameters, this approach significantly reduces training costs while effectively embedding specialized knowledge from the perovskite domain. As a result, the model internalizes and transfers domain knowledge relevant to specific materials science tasks. The fully trained Perovskite-R1 model thus retains robust general language modeling capabilities while acquiring expert-level proficiency in tasks such as materials structure design, defect regulation, and synthesis pathway recommendation.

To enable controlled application of the model in real-world scientific research tasks, we develop a structured prompt system tailored to the research workflow. Each prompt is divided into three components: task definition (e.g., Select suitable additives for defect passivation in FAPbI₃), scientific and application-oriented criteria (e.g., Prioritize heterocyclic compounds that can enhance device stability), and output format specification (e.g., Please provide the molecular structure and its anticipated mechanism of action). This prompt system supports both single-turn and iterative multi-turn interactions, allowing for interactive scientific dialogue with the model. Ultimately, under the guidance of materials chemistry experts, we design several experimental protocols based on the recommendations generated by Perovskite-R1 and carried out corresponding synthesis and validation in the laboratory. The results demonstrate that some of the novel additives suggested by the model exhibited excellent performance in defect passivation and stability enhancement, confirming their ability to provide innovative experimental insights for practical research applications.

Maintenance of reasoning ability

For the molecular library data, compounds are grouped into sets to ensure diversity and manageability. The o1 model is then tasked with interpreting the properties of each compound within a group using CoT-style reasoning, and subsequently identifying which molecules could serve as promising precursor additives. The resulting question–answer data from both the literature analysis and molecular library are merged to form a comprehensive fine-tuning dataset for QwQ-32B. Through this workflow, we derive the specialized Perovskite-R1 model.

This carefully orchestrated process ensures that the model’s advanced reasoning abilities are not only preserved but also enhanced in the context of the discovery of perovskite precursor additives. As illustrated in Figure 2, Perovskite-R1 is thus capable of synthesizing complex scientific knowledge and generating actionable, high-quality recommendations for experimental research.

Construction of prompts

The effectiveness of LLM outputs is highly dependent on the quality and design of the prompts provided during inference. Recognizing this crucial factor, we adopt a carefully crafted prompt strategy in our work. Our prompts are methodically structured to ensure clarity, precision, and contextual relevance, while also encompassing a broad spectrum of domain-specific expertise. By systematically incorporating task definitions, scientific criteria, and explicit output requirements, our approach not only guides the LLM toward generating accurate and insightful responses but also enhances its ability to address complex, specialized research questions with greater reliability and



Question:

Core Design Goals:

1. Defect passivation.
2. Control crystallization kinetics and film morphology.
3. Optimize energy-level alignment.....

Chain of Thought:

1. **Analysis of each compound**:
 - **Dimethylphosphine oxide**:
 - **Solubility**: It is likely to be soluble in DMF/DMSO as it has a phosphorus - oxygen bond which can interact with the polar solvents.....



Answer:

Dimethylphosphine oxide (CAS: 7211-39-4), 5-Hydroxy-2-Methyl-Benzoic Acid (CAS: 578-22-3), -Methoxy-4-nitrophenol (CAS: 3251-56-7) are selected as they are more likely to meet the criteria compared to others, with their functional groups potentially influencing

Figure 2: An example of dialogue. The user provides the prompt, then Perovskite-R1 first gives its thought process and following its final result.

depth. The prompt we design consists of three clearly delineated sections:

1. **Task Definition.** In the first section, we explicitly state the overarching goal: guiding the model to recognize and select appropriate chemical precursor additives for perovskite synthesis. For instance, the core mission of Perovskite-R1 can be conveyed through the formulation of specific objectives or requirements to be accomplished within this section.
2. **Objective Criteria.** The second section lays out the scientific and application-driven requirements. In this section, the focus is on distinguishing Perovskite-R1 from a standpoint that determines the suitability of additives or the specific requirements for the desired additives. This standpoint encompasses the reaction process that the additives are expected to fulfil and the properties they are required to exhibit, to the greatest extent possible.
3. **Expected Output Specification.** In the final section, we detail exactly what the model's response should include. As a case in point, the format and amount of perovskite-R1 output content can be specified, along with an analysis of the relevant choices.

In addition to the three primary points of the appeal, it is possible to add some additional requirements that may be valid. These could include the requirement for Perovskite-R1 to carefully analyse each additive, or the requirement for a suitable additive to not be made without careful consideration.

Moreover, given the extensive collection of drug libraries, it is imperative to ensure that perovskite-R1 has thoroughly evaluated all available drugs. To this purpose, a systematic approach involving traversal, screening, and subsequent iteration is adopted. The procedure entails the initial classification of all drugs into equal numbers of distinct groups. Thereafter, up to three additives from each

group are selected based on their suitability. These additives are then subjected to a subsequent screening step, wherein they are evaluated within their respective groups. This process is repeated until a satisfactory number of additives are identified. This strategy ensures that our drug library is fully utilised and demonstrates the superiority of LLM in screening judgment, as this task is difficult to accomplish in a short period of time by human labour, whereas Perovskite-R1 may be able to do it in a few hours or even minutes.

Experiment of Perovskite-R1

To rigorously evaluate the effectiveness of our model and enable fair comparison with other state-of-the-art language models, we constructed a comprehensive benchmark dataset focused on perovskite research. This dataset comprises a wide range of carefully curated question–answer pairs that span various domains of perovskite science. Each question was independently assessed by domain experts and categorized into three levels of difficulty—easy, medium, and hard—based on the complexity of the scientific concepts and the reasoning required.

We subsequently benchmark Perovskite-R1 against several prominent language models, including DeepSeek-R1, Gemini-2.5-Flash-Thinking, and others, using this dataset as a standardized evaluation platform. As summarized in Table 1, Perovskite-R1 consistently outperforms the competing models across all difficulty categories, with particularly excelling in challenging, domain-specific tasks. Notably, Perovskite-R1 has achieved a 10% enhancement in accuracy over the original QwQ-32B. These results provide compelling evidence for the superior capability and domain adaptation of Perovskite-R1 in addressing perovskite-related scientific problems, thereby substantiating both the efficacy and practical relevance of our approach.

Model	Benchmark		
	Easy	Medium	Hard
QwQ-32B	76.27	76.84	74.75
Doubao-seed-1.6-thinking	71.30	71.46	70.77
DeepSeek-R1	76.42	74.88	73.97
Gemini-2.5-flash-thinking	78.71	77.52	76.05
Perovskite-R1	86.92	85.06	84.63

Table 1: The performance of the Perovskite-R1 on benchmark dataset.

Discovery of suitable precursor additives

A primary objective in the development of Perovskite-R1 is to enable the identification of high-quality precursor additives for the synthesis of advanced perovskite materials. This section demonstrates the practical application of Perovskite-R1 in guiding the discovery and screening for precursor additives.

Utilizing carefully designed prompts that integrate multiple evaluation criteria, we conduct a comprehensive screening process with Perovskite-R1. The model efficiently generated a targeted shortlist of small-molecule additives that satisfy the specified requirements and are predicted to enhance both device efficiency and operational stability. These high-potential candidates are currently undergoing systematic experimental validation. This involves formulating precursor solutions, characterizing film morphology and optoelectronic properties, and fabricating complete solar cell devices to rigorously assess the actual impact of each additive on PCE, film quality, and long-term device

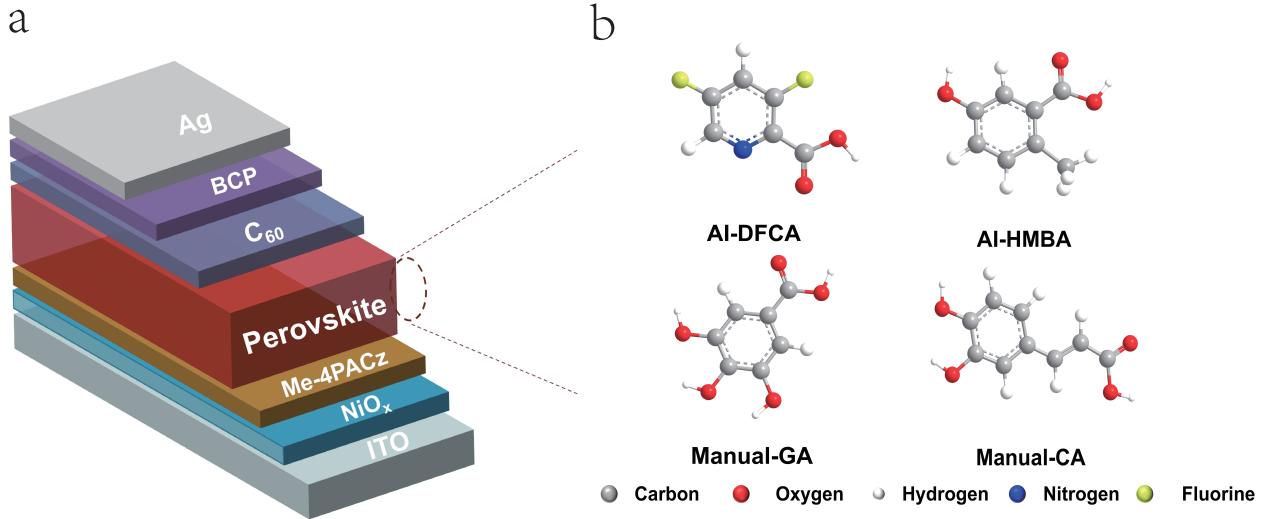


Figure 3: Device architecture and the selected molecules. a. Schematic of the PSC structure. b. The four precursor additives selected for analysis are DFCA and HMBA, which were chosen by Perovskite-R1, and GA and CA, which were chosen manually.

performance. This closed-loop framework highlights Perovskite-R1’s effectiveness in accelerating the discovery and evaluation of promising perovskite precursor additives.

To empirically validate the model’s predictions, a comparative experiment was conducted. Two types of additives were selected for systematic comparison. The first includes 3, 5-difluoropyridine-2-carboxylic acid (AI-DFCA) and 5-hydroxy-2-methylbenzoic acid (AI-HMBA), which were autonomously recommended by Perovskite-R1. The second comprises gallic acid (Manual-GA) and caffeic acid (Manual-CA), manually selected by researchers from the same molecular library based on domain knowledge, existing literature reports and experience. The chemical structure of all four additives is shown in Figure 3.

To ensure experiment accuracy, all four additives were introduced into the precursor solution of the perovskite component $\text{Cs}_{0.05}\text{MA}_{0.1}\text{FA}_{0.85}\text{PbI}_3$ at the same concentration (0.1 mg/mL), and the photovoltaic devices were fabricated using procedures, as shown in Figure 3. PCE measurements were conducted under standard conditions (AM 1.5G, 100 mW/cm²).

Figure 4 presents the J - V curves of the champion devices with and without the additive treatment. The experimental results in Table 2 show a clear and highly consistent trend. Both DFCA and HMBA, the additives selected by Perovskite-R1, enhance the device performance, achieving PCEs of 18.58% and 18.67%, respectively. This result directly validates the model’s ability to identify effective additives from a vast number of candidate molecules. In sharp contrast, the manually selected additives Manual-GA and Manual-CA demonstrate detrimental effects on device performance, reducing the PCE to 11.56% and 13.47%, respectively, substantially below the baseline value of 18.30% observed in the control devices. This contrast underscores the limitations of manual selection, even when guided by expert knowledge. It further emphasizes that cognitive biases and gaps in information coverage can significantly hinder additive screening in complex material systems.

Importantly, since all four experiments followed identical device fabrication processes and test conditions, the observed differences in efficiency can be directly attributed to the additive selection strategy. This set of data presents a strict opposing pattern: the two additives recommended by Perovskite-R1 both significantly improve the device efficiency, while the two additives selected manually both show inhibitory effects. This direct comparison reinforces that even expert-driven

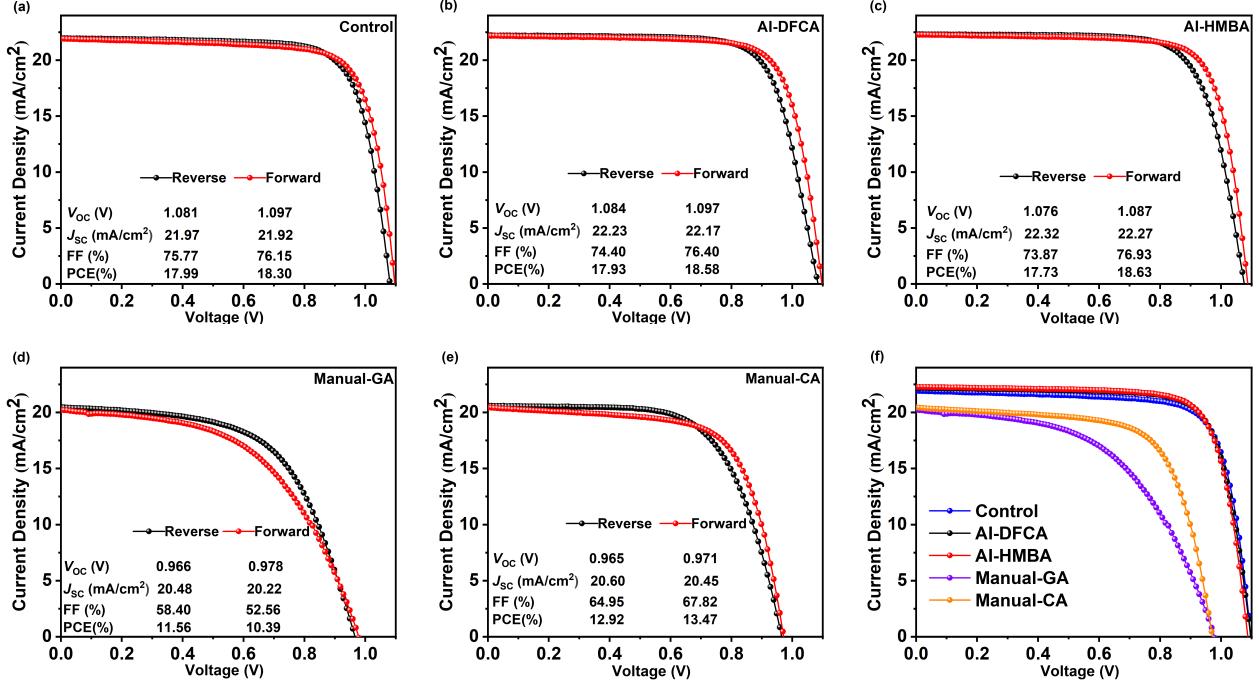


Figure 4: Current density-voltage (J - V) characteristic curves (forward and reverse scans) of PSCs. As illustrated in (a), the control group is represented, while (b) through (e) present the experimental outcomes for the four selected molecules. And (f) offers a comprehensive summary of the experimental results. Test conditions: scan rate of 10mV/10ms, AM1.5G illumination (100 mW/cm²). The open-circuit voltage (V_{OC}), short-circuit current density (J_{SC}), fill factor (FF), and PCE for both forward and reverse scans are labeled in the figure.

manual selection may fail to predict optimal candidates, whereas Perovskite-R1 has the ability to identify effective additives from a vast number of candidate molecules.

These findings provide strong evidence that Perovskite-R1 developed in this study significantly outperforms the manual screening method that relies on the subjective experience of researchers in the initial screening efficiency of perovskite precursor solution additives. The ability of the model to correctly predict the performance direction of additives demonstrates its high reliability and practical relevance.

The success of Perovskite-R1 stems from its in-depth learning of the structure-activity relationship between molecular features and additive functions, based on insights extracted from 1,232 professional and 33,269 molecular features. This comprehensive training enables the model to break through the limitations of individual cognition and the blindness of the traditional trial-and-error method, significantly improving the screening efficiency and reducing development cost. Overall, this result confirms the huge potential of AI-based models in solving complex material screening problems and accelerating the development of perovskite photovoltaic materials, providing a new approach for breaking through the traditional scientific research model.

Discussion

In this work, we address the inefficiency and reliance on expert intuition in designing defect compensation strategies for PSC precursor additives by introducing Perovskite-R1, a domain-specific LLM with advanced reasoning capabilities. We construct a high-quality instruction-tuning dataset comprising 1,232 scientific publications and 33,269 candidate materials, and fine-tune the QwQ-32B

Additives	V_{OC} (V)	J_{SC} (mA/cm 2)	FF (%)	PCE (%)
Control	1.097	21.92	76.15	18.30
Manual-GA	0.966	20.48	58.40	11.56
Manual-CA	0.971	20.45	67.82	13.47
AI-DFCA	1.097	22.17	76.40	18.58
AI-HMBA	1.087	22.27	76.93	18.63

Table 2: The photovoltaic parameters for PSCs with different additives.

base model to develop Perovskite-R1. Unlike conventional static prediction models, Perovskite-R1 systematically integrates domain knowledge and up-to-date experimental findings to generate innovative and practical material design solutions. Experimental validation confirms the model’s effectiveness, with several proposed strategies significantly enhancing material stability and device performance. This work not only advances the intelligent design of PSC materials but also provides a transferable framework for building high-performing domain-specific language models in other scientific fields.

Although Perovskite-R1 can accurately recognize terminology and provide reasonable answers in single-turn question answering, the depth of its output is limited by the phrasing of the question, often remaining superficial. For example, when addressing complex mechanisms such as defect passivation, the model tends to offer concise summaries without delving into detailed explanations. To address this issue, a preliminary solution is to introduce a multi-turn dialogue mechanism, enabling the model to gradually reason and elaborate through continuous questioning and contextual review, thus achieving more comprehensive and in-depth scientific analysis. In the future, further improvements can be made by incorporating reinforcement learning (such as RLHF) to optimize dialogue guidance strategies and building a multi-turn dialogue knowledge tracking module to enhance the model’s ability to understand consecutive questions. Besides, the control over structured experimental design tasks still has certain limitations. Currently, the model primarily generates experimental protocols through text prompts. However, parameter settings such as molar concentration, additive ratio, and spin-coating speed still require manual post-processing, as the model does not precisely model process feasibility or preparation-condition windows. As an initial solution, a structured prompt template has been developed, consisting of three main components: task definition, target criteria, and output format, which significantly improves the standardization of experimental recommendations. For future development, the introduction of a knowledge graph-based feasibility constraint checking module is planned, along with the integration of physical simulation models to provide parameter space mapping and automatic verification of boundary conditions.

Looking ahead, the development of Perovskite-R1 will focus on several key directions, including expanding the coverage of material systems, enhancing multi-task capabilities, and improving the synthesizability and experimental closed-loop capability of generated solutions. The current model centers on the design of defect passivation strategies for perovskite precursor additives, and has already demonstrated both intelligence and practicality in small-molecule additive screening and experimental validation. In the future, we aim to broaden the application boundaries of the model to encompass a wider range of tasks, such as interface engineering design, solvent system optimization, stability regulation strategies, and the co-design of device architectures. Leveraging Perovskite-R1’s strengths in chemical structure reasoning and condition recommendation, its applications can also be extended to the intelligent development of low-dimensional perovskites, perovskite-silicon tandem solar cells, and stable wide-bandgap material systems.

More importantly, Perovskite-R1 establishes an efficient closed-loop framework that integrates

literature mining, knowledge consolidation, solution generation, and experimental validation. This paradigm can be further generalized to other functional material fields, especially in the design of emerging energy storage materials (such as solid-state electrolytes and magnesium-based hydrogen storage materials), catalytic materials (for example, CO₂ reduction and electrocatalytic oxygen evolution), and flexible electronic materials (including polymers with strain-responsiveness and self-healing properties). By integrating with automated synthesis platforms and high-throughput experimental systems, Perovskite-R1 can further evolve into a “research assistant-type intelligent agent” that actively proposes hypotheses, validates feedback, and self-optimizes, thus propelling materials science from a human-centered, experience-driven exploration model to a model-centered, autonomous innovation paradigm.

Methods

Large language model For our baseline, we select QwQ-32B, a state-of-the-art causal language model based on the Qwen2.5-32B architecture, featuring approximately 32.5 billion parameters, of which 31 billion are dedicated to non-embedding computations [34]. The model is composed of 64 Transformer layers and incorporates a range of advanced architectural features, including Rotary Position Embedding (RoPE) for improved positional encoding, SwiGLU activation functions for enhanced non-linearity, RMSNorm for stable and efficient normalization, and Attention QKV bias for fine-tuning attention distributions.

One of the distinguishing aspects of QwQ-32B is its adoption of grouped-query attention (GQA), which utilizes 40 query heads partitioned into eight shared key–value heads. This design enables the model to achieve efficient and scalable sparse attention, making it capable of processing exceptionally long sequences—up to 131,072 tokens—without compromising computational efficiency. To adapt QwQ-32B for domain-specific tasks, we perform instruction tuning (supervised fine-tuning) using our curated perovskite instruction dataset, thereby endowing the model with specialized knowledge and reasoning abilities tailored to the discovery and design of PSC precursors.

The construction of dataset To construct a comprehensive and domain-specific training corpus, we curate a total of 1,232 research articles alongside a library of 33,269 candidate drug compounds relevant to perovskite precursor design. Due to the model’s fixed input token limitations, each research article is systematically segmented into passages of uniform length, yielding approximately 5,000 discrete text fragments for subsequent processing. In parallel, all candidate compounds—systematically identified by both their common names and CAS registry numbers—are classified into seven major therapeutic categories, reflecting the breadth of chemical diversity considered. This organizational strategy result in 4,753 individual data records representing alternative chemical agents. Together, these structured resources form a robust foundation for high-quality instruction tuning, ensuring that the model is exposed to a wide variety of scientific contexts and molecular candidates during training.

Chain-of-Thought generation For each text fragment and drug record, we leverage an LLM to automatically generate paired question–answer instances, each accompanied by detailed CoT reasoning. In the case of the research article data, each segmented passage is used as the input prompt; the model-generated analytical reasoning process is captured as the CoT, while the synthesized summary served as the final answer output. For the drug-related data, the compound names are provided as the primary input, prompting the model to produce a step-by-step analysis

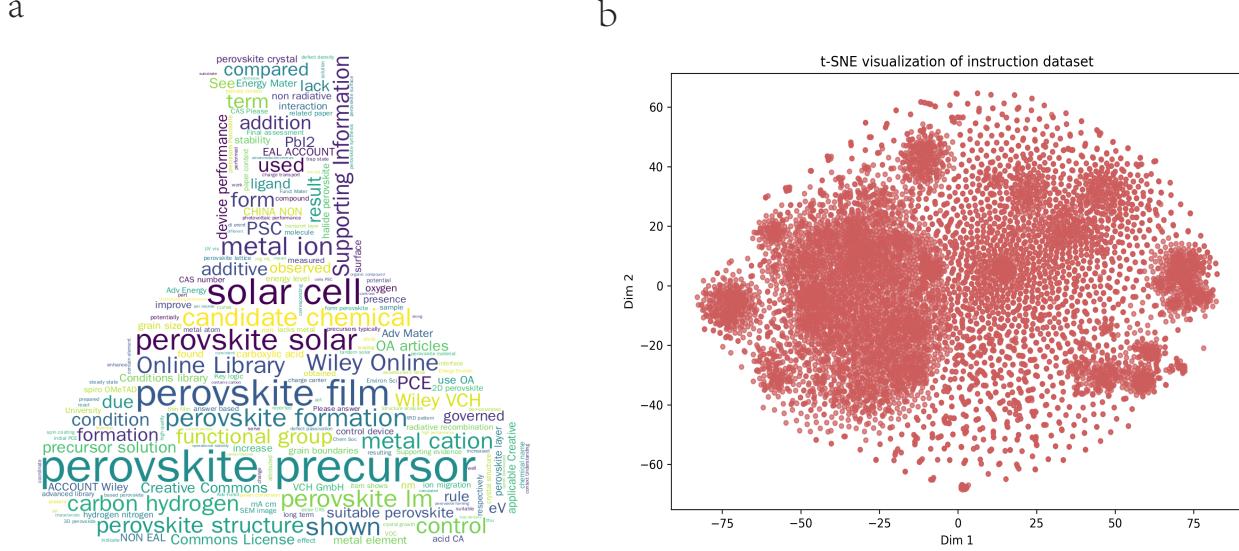


Figure 5: The word cloud and t-SNE visualization of the instruction dataset. a. The word cloud illustrates the terms that frequently appear in the instruction dataset, including perovskite, precursor and solar, all of which are relevant to the task at the core of this study. b. Two-dimensional t-SNE map of concatenated input–output TF-IDF vectors, using perplexity 30, learning rate 200, PCA initialization, and 1,000 iterations. Each point represents a data sample; spatial proximity reflects semantic similarity.

of each drug’s properties and mechanisms, followed by an informed assessment of its potential as a perovskite precursor additive as the concluding output.

This systematic approach results in a richly annotated CoT dataset that not only preserves the complexity and depth of scientific reasoning but also actively exercises and reinforces the model’s advanced reasoning capabilities during subsequent fine-tuning and evaluation. By structuring the data in this way, we ensure that the model is trained to deliver interpretable, logically sound, and contextually relevant responses across a diverse range of perovskite research scenarios.

Model fine-tuning During the instruction tuning stage, we adopt the QwQ-32B model as our foundation and implement parameter-efficient adaptation using the Low-Rank Adaptation (LoRA) technique across all weight matrices. For optimal performance, LoRA is configured with a rank of 16, an alpha scaling factor of 32, and a dropout rate of 0.1 to mitigate overfitting. Training is conducted using bfloat16 numerical precision, with a per-device batch size of 1, and gradients are accumulated over 8 steps to yield an effective batch size of 8. The learning rate is set to an initial value of 1e-4 and is decayed according to a cosine annealing schedule, incorporating a 5% warmup phase to stabilize early training dynamics. The model is trained for a total of 10 epochs. To further enhance computational efficiency, we utilize FlashAttention2, which enables rapid and memory-efficient attention calculations. This carefully orchestrated configuration ensures both training stability and the effective integration of domain-specific knowledge into the fine-tuned model.

Data availability

An example of the training set is available from <https://huggingface.co/datasets/JH976/Perovskite-R1>. The QwQ-32B model is publicly available through the Qwen organization’s model hub: <https://huggingface.co/qQwen/QwQ-32B>, released under Apache License 2.0.

Code availability

The code for splitting the paper to generate the dataset and for calling OpenAI o1 model is available at <https://huggingface.co/datasets/JH976/Perovskite-R1>. The code for fine-tuning the QwQ-32B model is based on the open-source LLaMA-Factory framework [52]. It includes all training scripts, configuration files (e.g., LoRA settings, optimizer, dataset pipelines), and inference utilities. The complete repository is publicly available at: <https://github.com/hiyouga/LLaMA-Factory>, released under Apache License 2.0.

References

- [1] Akihiro Kojima, Kenjiro Teshima, Yasuo Shirai, and Tsutomu Miyasaka. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *Journal of the american chemical society*, 131(17):6050–6051, 2009.
- [2] National Renewable Energy Laboratory. Interactive Best Research-Cell Efficiency Chart, June 2025. Accessed 2025-06-22.
- [3] Ze-Kai Bian, Zhenhuang Su, Yan-Hui Lou, Jing Chen, Run-Jun Jin, Chun-Hao Chen, Yu Xia, Lei Huang, Kai-Li Wang, Xingyu Gao, et al. Removal of residual additive enabling perfect crystallization of photovoltaic perovskites. *Angewandte Chemie International Edition*, 64(4):e202416887, 2025.
- [4] Liuwen Tian, Enbing Bi, Ilhan Yavuz, Caner Deger, Yuan Tian, Jingjing Zhou, Shaochen Zhang, Qingqing Liu, Jiahui Shen, Libing Yao, et al. Divalent cation replacement strategy stabilizes wide-bandgap perovskite for Cu (In, Ga) Se₂ tandem solar cells. *Nature Photonics*, 19:479–485, 2025.
- [5] Ali Hassan, Zhijie Wang, Yeong Hwan Ahn, Muhammad Azam, Abbas Ahmad Khan, Umar Farooq, Muhammad Zubair, and Yu Cao. Recent defect passivation drifts and role of additive engineering in perovskite photovoltaics. *Nano Energy*, 101:107579, 2022.
- [6] Mohsin Afroz, Ratneshwar Kumar Ratnesh, Swapnil Srivastava, and Jay Singh. Perovskite solar cells: Progress, challenges, and future avenues to clean energy. *Solar Energy*, 287:113205, 2025.
- [7] Jiaqi Yang, Panayotis Manganaris, and Arun Mannodi-Kanakkithodi. Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm. *The Journal of Chemical Physics*, 160(6), 2024.
- [8] Junya Wang, Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Feature selection in machine learning for perovskite materials design and discovery. *Materials*, 16(8):3134, 2023.
- [9] Islam M Abdellah and Ahmed El-Shafei. A machine learning approach for in silico prediction of the photovoltaic properties of perovskite solar cells based on dopant-free hole-transport materials. *New Journal of Chemistry*, 48(44):18666–18682, 2024.
- [10] Xiang Liu, Penglei Sun, Shuyan Chen, Longhan Zhang, Peijie Dong, Huajie You, Yongqi Zhang, Chang Yan, Xiaowen Chu, and Tong-yi Zhang. Perovskite-llm: Knowledge-enhanced large language models for perovskite solar cell research. *arXiv preprint arXiv:2502.12669*, 2025.

- [11] Seongmin Kim, Joshua Schrier, and Yousung Jung. Explainable synthesizability prediction of inorganic crystal polymorphs using large language models. *Angewandte Chemie International Edition*, 64(19):e202423950, 2025.
- [12] Jiye Han, Keonwoo Park, Shaun Tan, Yana Vaynzof, Jingjing Xue, Eric Wei-Guang Diau, Moungi G Bawendi, Jin-Wook Lee, and Il Jeon. Perovskite solar cells. *Nature Reviews Methods Primers*, 5(1):3, 2025.
- [13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [14] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [16] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [17] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [19] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihaot Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [22] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.
- [23] Seonok Kim. Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation. *arXiv preprint arXiv:2502.03004*, 2025.
- [24] Linqing Chen, Weilei Wang, Zilong Bai, Peng Xu, Yan Fang, Jie Fang, Wentao Wu, Lizhi Zhou, Ruiji Zhang, Yubin Xia, et al. Pharmagpt: Domain-specific large language models for bio-pharmaceutical and chemistry. *arXiv preprint arXiv:2406.18045*, 2024.

- [25] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):1–16, 2024.
- [26] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.
- [27] Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.
- [28] Vaibhav Mishra, Somaditya Singh, Dhruv Ahlawat, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, NM Krishnan, et al. Foundational large language models for materials research. *arXiv preprint arXiv:2412.09560*, 2024.
- [29] Zheyang Qu, Lu Yin, Zitong Yu, Wenbo Wang, et al. Coursegpt-zh: An educational large language model based on knowledge distillation incorporating prompt optimization. *arXiv preprint arXiv:2405.04781*, 2024.
- [30] Zhoumingju Jiang and Mengjun Jiang. Beyond answers: Large language model-powered tutoring system in physics education for deep learning and precise understanding. *arXiv preprint arXiv:2406.10934*, 2024.
- [31] Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*, 2023.
- [32] Cehao Yang, Chengjin Xu, and Yiyan Qi. Financial knowledge large language model. *arXiv preprint arXiv:2407.00365*, 2024.
- [33] Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, et al. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*, 2025.
- [34] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.
- [35] Xiao-Qi Han, Xin-De Wang, Meng-Yuan Xu, Zhen Feng, Bo-Wen Yao, Peng-Jie Guo, Ze-Feng Gao, and Zhong-Yi Lu. Ai-driven inverse design of materials: Past, present and future. *Chinese Physics Letters*, 2024.
- [36] Zhixun Li, Bin Cao, Rui Jiao, Liang Wang, Ding Wang, Yang Liu, Dingshuo Chen, Jia Li, Qiang Liu, Yu Rong, et al. Materials generation in the era of artificial intelligence: A comprehensive survey. *arXiv preprint arXiv:2505.16379*, 2025.
- [37] Mouyang Cheng, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. Ai-driven materials design: a mini-review. *arXiv preprint arXiv:2502.02905*, 2025.
- [38] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.
- [39] Yomn Alkabakibi, Congwei Xie, and Artem R Oganov. Graph neural network prediction of nonlinear optical properties. *arXiv preprint arXiv:2504.19987*, 2025.

- [40] Harikrishnan Sibi, Jovita Biju, and Chandra Chowdhury. Advancing 2d material predictions: superior work function estimation with atomistic line graph neural networks. *RSC advances*, 14(51):38070–38078, 2024.
- [41] Joshua Ojih, Mohammed Al-Fahdi, Yagang Yao, Jianjun Hu, and Ming Hu. Graph theory and graph neural network assisted high-throughput crystal structure prediction and screening for energy conversion and storage. *Journal of Materials Chemistry A*, 12(14):8502–8515, 2024.
- [42] Ramya Gurunathan, Kamal Choudhary, and Francesca Tavazza. Rapid prediction of phonon structure and properties using the atomistic line graph neural network (alignnn). *Physical Review Materials*, 7(2):023803, 2023.
- [43] Zijian Du, Luozhijie Jin, Le Shu, Yan Cen, Yuanfeng Xu, Yongfeng Mei, and Hao Zhang. Ctgnn: Crystal transformer graph neural network for crystal material property prediction. *arXiv preprint arXiv:2405.11502*, 2024.
- [44] Mingjian Wen, Matthew K Horton, Jason M Munro, Patrick Huck, and Kristin A Persson. An equivariant graph neural network for the elasticity tensors of all seven crystal systems. *Digital Discovery*, 3(5):869–882, 2024.
- [45] Faiza Qayyum, Murad Ali Khan, Do-Hyeun Kim, Hyunseok Ko, and Ga-Ae Ryu. Explainable ai for material property prediction based on energy cloud: a shapley-driven approach. *Materials*, 16(23):7322, 2023.
- [46] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.
- [47] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.
- [48] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.
- [49] Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steve Farrell, Benjamin Erichson, Michael W Mahoney, Andy Nonaka, and Zhi Yao. Matterchat: A multi-modal llm for material science. *arXiv preprint arXiv:2502.13107*, 2025.
- [50] Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.
- [51] Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. Llamp: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation. *arXiv preprint arXiv:2401.17244*, 2024.
- [52] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Acknowledgement: The work is supported by the National Natural Science Foundation of China (No.62476278, No.11934020). Computational resources have been provided by the Physical Laboratory of High Performance Computing at Renmin University of China.

Author contributions: Z.F.G., C.M., and Z.Y.L. contributed to the ideation and design of the research; X.D.W., Z.R.C., and Z.F.G. performed the research; X.D.W., Z.R.C., Z.F.G., C.M., and Z.Y.L. wrote and edited the paper; all authors contributed to the research discussions.

Corresponding authors: Correspondence and requests for materials should be addressed to Ze-Feng Gao (zfgao@ruc.edu.cn), Cheng Mu (cmu@ruc.edu.cn), and Zhong-Yi Lu (zlu@ruc.edu.cn).

Competing interests: The authors declare no competing interests.

Supplementary information: The supplementary information is attached.

A An example of prompt

The following is an example of the prompt utilized in the experiment. The “Core Design Goal” section aligns closely with the previously discussed “Task Definition” section, ensuring consistency between the design’s foundational objectives and the overall task framework. Furthermore, the “Key Design Principles & Strategies” and “Critical Design Considerations” sections correspond directly to the “Objective Criteria” section, reflecting the essential priorities and strategic approaches established earlier. Finally, the “Output Requirements” and “Additional Filtering Rule” sections directly map to the “Expected Output Specification”, providing a clear link between the intended results and the defined output standards. In addition to the aforementioned three sections, the prompt contains several supplementary components that are deemed essential for the comprehensive understanding of the given subject or task.

Core Design Goals:

1. Defect passivation.
2. Control crystallization kinetics and film morphology.
3. Optimize energy-level alignment.

Key Design Principles & Strategies:

1. Lewis acid-base interactions.
2. Crystallization modulators.
3. Multifunctional additive design.

Critical Design Considerations:

1. Solubility in DMF/DMSO.
2. Compatibility with perovskite chemistry and subsequent layers.
3. Prefer commercially available or natural/food-grade compounds.

Output Requirements:

- Analyze functional groups and mechanistic role.
- Evaluate processing compatibility.
- Rank 3–5 top candidate additives by expected PV efficiency improvement.
- For each: chemical name + CAS number; rationale/mechanism; primary function(s).

Additional Filtering Rule:

- **If fewer than 3 candidates truly meet ALL the core goals and critical considerations, do not force the output to contain 3–5 items. Instead, only list those compounds that genuinely satisfy the criteria.**

Here is a list of candidate compounds from our chemical database:

- Octa(aminophenyl)-T8-silesquioxane (CAS: 518359-82-5)
- 4-Cyanopyridine (CAS: 100-48-1)
- Quinuclidine (CAS: 100-76-5)
- N-Nitrosopiperidine (CAS: 100-75-4)
- potassium phenolate (CAS: 100-67-4)
- 3-Pyridinemethanol (CAS: 100-55-0)
- N-Benzyl-N,N-dimethyl-1-phenylmethanaminium chloride (CAS: 100-94-7)
- 3-Hydroxybenzaldehyde (CAS: 100-83-4)

- Butane, 1-(1,1-dimethylethoxy)- (CAS: 1000-63-1)
- 1H-Pyrrolo[2,3-b]pyridine-3-carboxaldehyde, 4-bromo- (CAS: 1000340-35-1)
- 1H-Pyrrolo[3,2-b]pyridine, 5-bromo- (CAS: 1000341-51-4)
- Benzoic acid, 3-amino-5-bromo-2-methyl-, methyl ester (CAS: 1000342-11-9)
- 6-Chloro-1H,2H,3H-pyrrolo[3,2-c]pyridin-2-one (CAS: 1000342-80-2)
- 1H-Pyrrolo[3,2-c]pyridine, 6-bromo- (CAS: 1000342-71-1)
- 1H-Indazole, 6-bromo-5-methyl- (CAS: 1000343-69-0)
- 3-Benzofuranacetic acid, 2,3-dihydro-6-hydroxy- (CAS: 1000414-37-8)
- 3-Pyridazinecarboxylic acid, 1,6-dihydro-1-methyl-6-oxo- (CAS: 100047-66-3)
- 2-Pyridineacetic acid, 5-fluoro- (CAS: 1000515-83-2)
- 1H-Indazole, 7-bromo-1-methyl- (CAS: 1000576-59-9)
- Benzonitrile, 3-bromo-4-iodo- (CAS: 1000577-94-5)

Please screen and rank these compounds according to the prompt above.

****Final Output Instructions:****

- ****Output only the selected additives, each enclosed in a single set of braces .****
- ****Within each brace, include only the chemical name and CAS number.****
- Do not output any extra text or explanation beyond these braces.
- Ensure that your final selections are fully consistent with your internal reasoning.

B An example of output

The following example illustrates the potential efficacy of Perovskite-R1. In light of the example, we have elected to utilize 5-hydroxy-2-methylbenzoic acid, *i.e.* AI-HMBA, as the additive for the ensuing experiment, with the objective of identifying an optimal perovskite additive. The segment between `<think>` and `</think>` represents the chain of thought, which can be regarded as the reasoning process of Perovskite-R1. Meanwhile the segment following `</think>` is the final output of the input.

```
<think>
1. **Analysis of each compound**:
   - **Dimethylphosphine oxide**:
     - **Solubility**: It is likely to be soluble in DMF/DMSO as it has a phosphorus - oxygen bond which can interact with the polar solvents. It has a Lewis basic phosphorus atom that could potentially participate in Lewis acid - base interactions with perovskite precursors. It could act as a ligand or influence the coordination environment during perovskite formation, thus having the potential to control crystallization kinetics.
     - **Defect passivation**: The lone pair on the phosphorus atom might be able to interact with defects in the perovskite lattice, potentially passivating them.
     - **Energy - level alignment**: It may influence the energy levels through its interaction with the perovskite structure, but more detailed analysis would be needed.
   - **5 - Hydroxy - 2 - Methyl - Benzoic Acid**:
     - **Solubility**: It has a carboxylic acid group which can hydrogen bond with the solvent molecules in DMF/DMSO, so it is likely to be soluble.
     - **Lewis acid - base interactions**: The carboxylic acid group is a Lewis acid, and it could interact with Lewis basic sites in the perovskite precursor, influencing the crystallization process.
```

- **Defect passivation**: The hydroxyl group and the carboxylic acid group could potentially interact with defects in the perovskite, but the extent and mechanism would require further study.

- **Cyano - 3-(3 - chlorophenylethyl)pyridine**: - **Solubility**: The presence of the cyano group and the aromatic rings may make it less soluble in DMF/DMSO compared to some other compounds. The cyano group is a weak Lewis acid, and the pyridine ring has a nitrogen atom with lone pairs that could participate in Lewis acid - base interactions. However, the overall solubility might be a concern.

- **Crystallization control**: The Lewis acid - base interactions could potentially influence the growth of perovskite crystals, but the solubility issue needs to be addressed first.

- **Defect passivation**: The cyano and pyridine groups might interact with defects, but again, solubility is a limiting factor.

- **(1R,4R)-Methyl 4 - formylcyclohexanecarboxylate**: - **Solubility**: It has a carbonyl group which can interact with the solvent through hydrogen bonding, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The carbonyl group can act as a Lewis acid, and the ester group can participate in some interactions. It could potentially influence the coordination chemistry during perovskite formation.

- **Defect passivation**: The functional groups might interact with defects, but more research on the specific interactions would be needed.

- **Ethyl 3-(3 - nitrophenyl)-3 - oxopropanoate**: - **Solubility**: The nitro group and the carbonyl group can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The carbonyl group can act as a Lewis acid, and the nitro group may influence the electronic properties of the molecule. It could potentially affect the crystallization process.

- **Defect passivation**: The interactions might help passivate defects, but the mechanism would need to be explored further.

- **2,3,5 - Trichlorothiophene**:

- **Solubility**: It has a relatively non - polar structure, so it is less likely to be soluble in DMF/DMSO compared to compounds with more polar functional groups.

- **Lewis acid - base interactions**: The sulfur atom has some potential for weak Lewis acid - base interactions, but overall, due to its solubility limitations, it may not be a good candidate for perovskite precursor additives.

- **Defect passivation**: The interactions are likely to be weak due to the lack of strong polar functional groups.

- **L - Valine, N-(1 - oxopentyl)-N-[2'-(2H - tetrazol - 5 - yl)[1,1'-biphenyl]-4 - yl]methyl**:-

- **Solubility**: It has multiple functional groups like the amide, tetrazole, and aromatic rings. The amide and tetrazole groups can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The tetrazole group can act as a Lewis base, and the amide group can participate in some interactions. It could potentially influence the perovskite crystallization.

- **Defect passivation**: The molecule might interact with defects through its various functional groups.

- **1,2 - Pyrrolidinedicarboxylic acid, 4,4 - diMethyl-, 1-(1,1 - diMethylethyl) ester, (2R)-**:-

- **Solubility**: It has ester groups which can interact with the solvent through hydrogen bonding, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The ester groups can participate in some interactions, and the molecule could potentially influence the coordination environment during perovskite formation.

- **Defect passivation**: The interactions might help passivate defects, but more study on the specific mechanisms is required.

- **Benzoic acid, 4 - ethoxy-, ethyl ester**:

- **Solubility**: It has an ester group and an ethoxy group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The ester group can participate in some interactions, but it may not have a strong impact on perovskite crystallization compared to some other compounds with more reactive functional groups.

- **Defect passivation**: The interactions are likely to be weak for defect passivation.

- **NU6140**:

- **Solubility**: Without knowing its exact structure in detail, it's difficult to assess solubility. But if it has polar functional groups, it may be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: If it has appropriate functional groups like nitrogen or oxygen with lone pairs, it could participate in Lewis acid - base interactions.

- **Defect passivation**: It might interact with defects, but more information about its structure is needed.

- **Propanoic acid, 3 - butoxy-, butyl ester**:

- **Solubility**: It has an ester group and an alkoxy group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The ester group can participate in some interactions, but it may not have a strong influence on perovskite crystallization.

- **Defect passivation**: The interactions are likely to be weak for defect passivation.

- **tert - butyl N-(4 - bromo - 1,3 - thiazol - 2 - yl)carbamate**:

- **Solubility**: It has a carbamate group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The carbamate group can participate in some interactions, and the bromine atom may influence the electronic properties. It could potentially affect the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more research on the specific interactions is needed.

- **11 - Oxa - 2,7,9 - triazatridec - 7 - enoic acid, 3 - carboxy - 8 - [(1,1 - dimethylethoxy)carbonyl]amino - 12,12 - dimethyl - 10 - oxo-, 1-(9H - fluoren - 9 - ylmethyl) ester, (3S)-**:

- **Solubility**: It has multiple functional groups like carboxylic acid, carbamate, and aromatic rings. It is likely to be soluble in DMF/DMSO due to the presence of polar functional groups.

- **Lewis acid - base interactions**: The carboxylic acid and carbamate groups can participate in Lewis acid - base interactions, potentially influencing the perovskite crystallization.

- **Defect passivation**: The molecule might interact with defects through its various functional groups.

- **2 - Methoxy - 4 - nitrophenol**:

- **Solubility**: It has a hydroxyl group and a nitro group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The hydroxyl group can act as a Lewis base, and the nitro group can influence the electronic properties. It could potentially affect the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more study on the specific mechanisms is required.

- **Ethyl N - Boc - 3 - oxopyrrolidine - 2 - carboxylate**:

- **Solubility**: It has an ester group and a carbamate group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The carbamate group can participate in some interactions, and the Boc group may influence the electronic properties. It could potentially affect the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more research on the specific interactions is needed.

- **tert - Butyl 3 - cyano - 4 - oxopiperidine - 1 - carboxylate**:

- **Solubility**: It has an ester group and a cyano group which can interact with the solvent, so it is likely to be soluble in DMF/DMSO.

- **Lewis acid - base interactions**: The cyano group can act as a Lewis base, and the ester group can participate in some interactions. It could potentially influence the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more study on the specific mechanisms is required.

- **3 - Aminopiperidine - 2,6 - dione hydrochloride**:

- **Solubility**: The presence of the hydrochloride salt may increase its solubility in DMF/DMSO. The amine group can participate in Lewis acid - base interactions.

- **Lewis acid - base interactions**: The amine group can interact with Lewis acidic sites in the perovskite precursor, potentially influencing the crystallization process.

- **Defect passivation**: The interactions might help passivate defects, but more research on the specific interactions is needed.

- **2 - COUMARANONE**:

- **Solubility**: It has an aromatic structure with a carbonyl group. It may have some solubility in DMF/DMSO due to the carbonyl group's ability to interact with the solvent.

- **Lewis acid - base interactions**: The carbonyl group can act as a Lewis acid, and it could potentially interact with Lewis basic sites in the perovskite precursor.

- **Defect passivation**: The interactions might help passivate defects, but more study on the specific mechanisms is required.

- **4 - CHLORO - 6,7 - DIMETHOXY - QUINOLINE - 3 - CARBOXYLIC ACID ETHYL ESTER**:

- **Solubility**: It has an ester group and an aromatic structure with a chlorine atom. It is likely to be soluble in DMF/DMSO due to the ester group's interaction with the solvent.

- **Lewis acid - base interactions**: The carbonyl group in the ester can participate in some interactions, and the chlorine atom may influence the electronic properties. It could potentially affect the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more research on the specific interactions is needed.

- **1H - Indole - 3 - butanoic acid, monopotassium salt**:

- **Solubility**: The presence of the potassium salt may increase its solubility in DMF/DMSO. The carboxylic acid group can interact with the solvent.

- **Lewis acid - base interactions**: The carboxylic acid group can act as a Lewis acid, and the indole ring may participate in some interactions. It could potentially influence the perovskite crystallization.

- **Defect passivation**: The interactions might help passivate defects, but more study on the specific interactions is needed.

2. **Final assessment**: - **Top candidates**:

- Dimethylphosphine oxide (CAS: 7211 - 39 - 4): It has a phosphorus - oxygen bond that can interact with perovskite precursors through Lewis acid - base interactions, potentially controlling crystallization kinetics and passivating defects.

- 5 - Hydroxy - 2 - Methyl - Benzoic Acid (CAS: 578 - 22 - 3): The carboxylic acid and hydroxyl groups can participate in Lewis acid - base interactions, influencing crystallization and potentially passivating defects.

- 2 - Methoxy - 4 - nitrophenol (CAS: 3251 - 56 - 7): It has a hydroxyl group and a nitro group that can interact with the solvent and participate in Lewis acid - base interactions, potentially affecting crystallization and defect passivation.

- **Other compounds**:

- Compounds like (1R,4R)-Methyl 4 - formylcyclohexanecarboxylate, Ethyl 3-(3 - nitrophenyl)-3 - oxopropanoate, L - Valine, N-(1 - oxopentyl)-N-[[2'-(2H - tetrazol - 5 - yl)[1,1'-biphenyl]-4 - yl]methyl] -, 1,2 - Pyrrolidinedicarboxylic acid, 4,4 - diMethyl-, 1-(1,1 - diMethylethyl) ester, (2R)-, tert - butyl N-(4 - bromo - 1,3 - thiazol - 2 - yl)carbamate, 11 - Oxa - 2,7,9 - triazatridec - 7 - enoic acid, 3 - carboxy - 8 - [(1,1 - dimethylethoxy)carbonyl]amino]-12,12 - dimethyl - 10 - oxo-, 1-(9H - fluoren - 9 - ylmethyl) ester, (3S)-, Ethyl N - Boc - 3 - oxopyrrolidine - 2 - carboxylate, tert - Butyl 3 - cyano - 4 - oxopiperidine - 1 - carboxylate, 3 - Aminopiperidine - 2,6 - dione hydrochloride, 2 - COUMARANONE, 4 - CHLORO - 6,7 - DIMETHOXY - QUINOLINE - 3 - CARBOXYLIC ACID ETHYL ESTER, 1H - Indole - 3 - butanoic acid, monopotassium salt also show potential but have either solubility concerns or less obvious functional group - based mechanisms for perovskite additive functions compared to the top three candidates.

In conclusion, the top three candidates are Dimethylphosphine oxide, 5 - Hydroxy - 2 - Methyl - Benzoic Acid, and 2 - Methoxy - 4 - nitrophenol, which show promise in meeting the core design goals through their Lewis acid - base interactions and potential for influencing crystallization kinetics and defect passivation. The other compounds also have some potential but require further investigation and optimization of processing conditions to fully assess their suitability as perovskite precursor additives.

</think>
Dimethylphosphine oxide (CAS: 7211-39-4), 5-Hydroxy-2-Methyl-Benzoic Acid (CAS: 578-22-3), 2-Methoxy-4-nitrophenol (CAS: 3251-56-7) are selected as they are more likely to meet the criteria compared to others, with their functional groups potentially influencing crystallization, defect passivation, and energy-level alignment, and considering solubility and compatibility.