Hypothesis Generation for Materials Discovery and Design Using Goal-Driven and Constraint-Guided LLM Agents

Anonymous ACL submission

Abstract

Materials discovery and design are critical for advancing technology across various industries, enabling the development of materials tailored for specific applications. To accelerate this process, recent research has turned to leveraging Large Language Models (LLMs) for scientific discovery, particularly in materials science. Building on this trend, we explore how LLMs can be harnessed to generate viable hypotheses that, when evaluated and verified in the laboratory, can expedite materials discovery and design. To assess the potential of LLMs in this domain, we collaborated with materials science experts to curate a novel dataset derived from recent journal publications on materials discovery and design. This dataset contains detailed information about goals focused on developing materials for real-world applications, along with relevant constraints and the methods employed to achieve these goals. Leveraging this dataset, we developed a method using LLM-based agents that, when provided with a goal and its constraints, generate hypotheses regarding the methods needed to achieve that goal. To evaluate the relevance and quality of the generated hypotheses compared to the actual methods from the dataset, we introduce a scalable evaluation metric developed with material science experts and evaluate our agents. We believe that our curated dataset, proposed method, and evaluation metrics will facilitate future research efforts aimed at accelerating materials discovery and design using LLMs.

1 Introduction

The discovery and design of materials to meet specific application needs is essential to address critical challenges we face today (Jain et al., 2013). Traditional methods for materials discovery and design are time-intensive and resource-heavy, requiring researchers to conduct extensive literature reviews, explore vast compositional, chemical, and structural spaces through simulations, and perform laborious lab-based experiments (Davies et al., 2016; Hautier et al., 2012). These challenges often lead to lengthy development cycles, slowing innovation and delaying the deployment of new technologies. As demand for more efficient, durable, and sustainable materials rises, accelerating materials discovery and design has become critical to meet these growing needs.

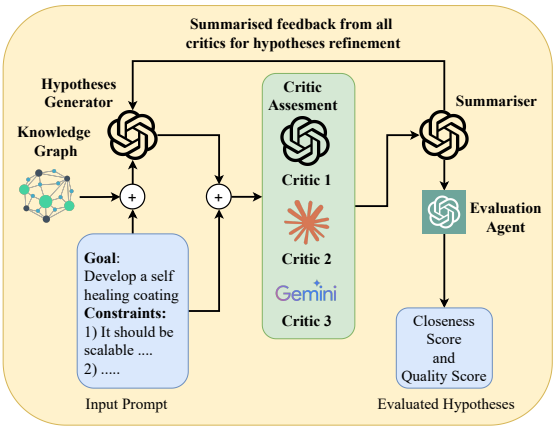

Figure 1: Given the Input prompt and Knowledge Graph, the Hypotheses Generator (GPT-4o) generates 20 hypotheses, evaluated by Critic 1 (GPT-4o), Critic 2 (Claude 3.5 Sonnet), and Critic 3 (Gemini 1.5 Flash). The summarizer consolidates the feedback. If the critics do not unanimously agree on all hypotheses, it sends them back for refinement; otherwise, the hypotheses proceed to the Evaluation Agent (GPT-4o1) for scoring

Recent advances in machine learning and data-driven approaches have accelerated materials discovery and design by enabling the prediction of material structures and properties(Liu et al., 2017; Oganov et al., 2019). Deep learning and generative models have shown promise in proposing novel materials (Chen and Ong, 2022; Merchant et al., 2023; Court et al., 2020; Xie et al., 2021; Ren et al., 2022; Fung et al., 2022; Long et al., 2021). However, these methods heavily depend on extensive training datasets and don't incorporate natural language inputs for predictions, limiting their accessibility

and flexibility in hypothesis generation. To address this limitation, recent research in natural language processing (NLP) and material sciences have begun investigating the application of LLMs for hypothesis generation in materials discovery and design. However, before applying LLMs in this field, it is essential to assess their ability to generate viable hypotheses specific to this domain. Existing work only assesses the domain understanding of LLMs by evaluating them on narrow tasks related to chemistry or material science(Zaki et al., 2024; Guo et al., 2023). Current methods for hypotheses generation in this domain that utilize LLMs focus on specific materials or properties and heavily rely on domain-specific tools (Jia et al., 2024; Sprueill et al., 2024; Ghafarollahi and Buehler, 2024). Our aim in this work is to explore how LLMs can generate potential hypotheses for real world materials discovery and design problems that can be tested in the lab, ultimately accelerating the materials design and discovery process. In practice, human researchers approach material development with specific goals in mind, utilizing constraints to guide the discovery process toward optimal solutions. This goal-oriented and constraint-driven methodology is essential for navigating complex design challenges and achieving desired material properties. This raises an intriguing question: can LLMs emulate the human approach to materials design by generating viable hypotheses when provided with specific goals and constraints? If so, LLMs could significantly accelerate the materials discovery process by proposing innovative solutions that can be evaluated and verified experimentally.

To investigate this possibility, we developed a novel dataset curated in collaboration with materials science experts called MatDesign. This dataset comprises 50 recently published papers from prominent journals on the ScienceDirect platform, carefully selected to evaluate LLMs' capabilities in hypothesis generation for real-world materials discovery and design goals. Each entry includes a goal derived from the respective paper and relevant constraints, serving as input prompts for LLMs to design or develop materials for specific applications. Additionally, the dataset provides brief descriptions of the materials and methods employed to achieve these goals. Building upon this dataset, we constructed an LLM-based agentic framework called AccelMat. Our framework generates hypotheses across a diverse range of real-world applications, incorporating various combinations of materials and properties without relying on domain-specific tools. It features an LLM-based hypothesis generator, a multi-LLM critic system that provides iterative feedback, and a summarizer agent that consolidates this feedback for hypothesis refinement. To evaluate the performance of our LLM-based system, we developed a scalable evaluation metric with expert input, based on nine key criteria assessing the closeness and quality of the generated hypotheses compared to the methods described in the dataset. This evaluation system utilizes GPT-4o1 to compare the generated hypotheses against the dataset's methods, calculating closeness and quality scores as defined by the nine criteria. Additionally, we conducted expert evaluations to provide a comprehensive assessment of the system's capabilities.

Dataset	MaScQA	ChemLLMBench	MatDesign
Constraints	×	×	✓
Real Problems	×	×	✓
No Data Leakage	×	×	✓
Difficulty Level	UnderGrad	Research	Research

Table 1: Comparison of existing datasets and benchmarks to our benchmark, MatDesign. Our benchmark provides constraints, has real-world goal-oriented problems, is not present in the pre-training of LLMs, and contains research-level problems.

2 Background and Related Work

2.1 LLMs for Materials Discovery and Design

In recent years, there has been a growing body of work exploring the use of Large Language Models (LLMs) in the field of materials discovery and design. These efforts can be broadly categorized into two main approaches. The first approach involves fine-tuning LLMs on domain-specific datasets with specific architectures (Sirumalla et al., 2024; Özçelik et al., 2024; Song et al., 2023) allowing the models to generate more accurate and relevant predictions for material properties (Jacobs et al., 2024) and potential applications. By tailoring the models to the language and concepts specific to materials science, researchers have been able to enhance the models' ability to generate hypotheses that are grounded in the domain's unique challenges. However, these approaches are resource intensive and dependent on large training datasets.

The second category includes single or multi-LLM agent-based frameworks, where LLMs are in-

tegrated with domain-specific tools and databases to actively participate in the materials discovery and design process. In these frameworks (Jia et al., 2024; Sprueill et al., 2024), LLMs are integrated with domain specific tools, and data repositories to generate and refine material hypotheses, combining AI's power with domain-specific knowledge. This approach creates a more autonomous system that iteratively refines material suggestions based on feedback from simulations, experiments, or added constraints. While promising, these methods are often limited to specific materials or properties, which restricts their generalizability. Additionally, they rely on domain-specific tools, which may not always be available for a wide range of tasks. Our approach addresses these limitations by investigating how well LLM-based agents, without the need for domain-specific tools, can generate hypotheses for a broad variety of applications.

LLM Agent	Range of Mat.	Range of Prop.	Tool Free
LLMatDesign	✓	×	✓
ChemReasoner	×	✓	×
SciAgents	×	✓	✓
ACCELMAT	✓	✓	✓

Table 2: Comparison of existing LLM agent frameworks with our framework, ACCELMAT, for materials design and discovery hypotheses generation. Our framework covers a wide range of materials and properties while being independent of domain-specific tools.

3 Methodology

In this section, we present the three core components of our approach: 1) a curated dataset specifically designed to evaluate materials discovery and design hypotheses generation capabilities of LLMs 2) an LLM-based multi-agent framework developed for hypotheses generation and refinement; and 3) a novel evaluation framework to assess the quality and closeness of generated hypotheses.

3.1 Novel Dataset Construction and Rationale

To evaluate the innovative capacity of LLMs on materials discovery and design hypotheses generation, we curated a specialized dataset. This collection comprises information extracted from 50 papers published or made available online after December 2023, sourced from prominent journals such as Nature, Nature Communications, and Progress in Organic Coatings. The dataset includes manually extracted and formulated elements for each paper:

- **Goal statements**, outlining the development of a material for a specific application. This forms first component of the input given to our multi-agent framework.
- **Key constraints** to guide the hypotheses generation process in the desired direction. This forms final component of the input given to our multi-agent framework.
- **Material names and compositions**: Describes the key materials and their compositions extracted from the paper. This forms the first component of our ground truth.
- **Methods**. Described the synthesis methods in brief extracted from the paper. This forms the final component of our ground truth.

Materials science domain experts assisted in all extractions and formulations, ensuring data accuracy and relevance. Importantly, we took measures to prevent any leakage of ground truth information (material names, compositions, and synthesis methods) in the goal statements and constraints. This precaution maintains evaluation integrity, ensuring generated hypotheses are genuinely novel and not influenced by inadvertently provided information

The selection of post-December 2023 publications is crucial, as it ensures that the material compositions, and synthesis methods that serve as ground truth hypotheses related to the respective goal statements are not present within the knowledge base of state-of-the-art closed-source LLMs used in our framework. These LLMs have a training data cutoff of December 2023. This temporal constraint allows us to assess the model's ability to generate novel hypotheses based on its internalized understanding of materials science principles, rather than merely retrieving existing information.

3.2 LLM based Multi-Agent Framework

3.2.1 Framework Architecture

The LLM-based multi-agent framework designed for this study consists of four key components, each playing a critical role in the iterative process of hypothesis generation and refinement.

1. **Hypotheses Generation Agent (HGA)**: The first component is the Hypotheses Generation

Goal	Constraints	Materials and Methods
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. The coating should autonomously recover its protective functionality after mechanical damage and initiate a self-healing process triggered by environmental factors. The method should address challenges related to the implementation of self-healing agents, while providing robust corrosion resistance and allowing for multiple healing cycles.	1) The material should incorporate a self-healing mechanism triggered by a single environmental factor (e.g., water) to facilitate immediate and efficient damage recovery in corrosion-prone environments. 2) The self-healing material should allow multiple healing events. 3) The coating must maintain its structural integrity and protective abilities even after mechanical damage, preventing corrosion from initiating at the damaged site.	Materials: Core-shell nanofibers synthesized using coaxial electrospinning. Organosilane compounds, specifically silyl esters, used as the self-healing agent. Metallic substrates (e.g., steel) for corrosion tests. Methods: Coaxial electrospinning of core-shell nanofibers with an organosilane compound (silyl ester) as the healing agent. Overcoming scalability challenges with spray painting: A viable spray-painting technique was implemented to apply the nanofiber-based coating, making it scalable for large structures. .....

Table 3: An instance of our dataset extracted from Spera et al. (2024)

Agent (HGA) (GPT-4o), which is responsible for generating multiple hypotheses, along with the reasoning behind each, based on the goal statement and constraints provided by a human expert from our curated dataset. The agent leverages the LLM's parametric knowledge to produce a variety of hypotheses that aim to achieve the goal while satisfying the provided constraints. The Hypotheses generation agent is initialised with a dynamically tailored system prompt and a user prompt and is tasked to generate hypotheses in a structured format. More details about the system prompt and user prompt in Appendix B.3.

2. **Critic Agents (CA)**: The second component involves critic agents based on three state-of-the-art closed-source LLMs (GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 flash). These agents are provided with the same goal statement and constraints given to the hypothesis generation agent, along with the generated hypotheses. Acting as critics, their task is to evaluate each suggestion from the hypothesis generation agent and provide feedback for future iterations. Each critic assesses whether the suggestions help achieve the goal and adhere to all constraints, offering detailed reasoning and comprehensive feedback to guide future hypothesis generation cycles. Like the hypothesis generation agent, these critic agents are initialized with specifically designed system and user prompts. More details can be found in the Appendix B.4

3. **Summarizer Agent(SA)**: The summarizer Agent's (GPT-4o) task is to consolidate the feedback given by all three critic LLMs, structure it, and then pass it on as comprehensive feedback to the hypotheses generator for refined generation. This agent is also initialised with specific system prompt. More details in Appendix B.5.

4. **Evaluation Agent (EA)**: The evaluation agent is used to evaluate the closeness and quality of the generated hypotheses.

The Hypothesis Generator Agent, Critic Agents, and Summarizer Agent form the hypothesis generation refinement framework. The Evaluation Agent, however, is not part of this framework; it is used solely to evaluate the generated hypotheses.

4 Experiments

We conducted experiments with our designed LLM based Multi-Agent Framework with three different configurations.

4.1 Hypotheses Generation without Feedback from Critics

In this setup (refer Appendix C.1 for the prompt), we employ a Hypothesis Generation Agent (HGA) alongside three Critic Agents (CAs). The HGA receives a goal statement and a set of constraints provided by a human expert, which are randomly sampled from our curated dataset. The HGA's task is to generate 20 hypotheses, each with detailed reasoning, in a structured format aimed at achieving the goal while satisfying the constraints simultaneously. The three CAs then independently evaluate the 20 generated hypotheses. Each CA reviews the hypotheses alongside the goal statement and constraints, assessing whether each hypothesis effectively addresses the goal and satisfies the constraints. Each CA returns the hypotheses that, in

Algorithm 1 Hypotheses Generation with Critic Feedback and Knowledge Graphs

Require: Goal Statement G_S, Constraint List C, Knowledge Graph K_G, Critic LLMs: $[C_1, C_2, C_3]$
Hypotheses Generator H_G, Summarizer S
Number of Cycles $max_cycles = 5$
$H_H \leftarrow []$ Hypothesis History initialized as empty list
$F_{AC} \leftarrow []$ Feedback from all critics initialized as empty list
Generate Initial Hypothesis:
$V_h \leftarrow H_G(G_S, C, K_G)$
$H_H.append(V_h)$
if *all* V_h *agreed by* C_1, C_2, C_3 **then**
  return V_h
end
else
  for $cycle = 1$ *to* max_cycles **do**
    $F_{AC} \leftarrow feedback(G_S, C, H_H[-1])$ Feedback from all critics
    $S_F \leftarrow summariser(F_{AC})$ Summarized Feedback from all critics
    $R_h \leftarrow H_G(G_S, C, K_G, H_H[-1], S_F)$ Refined Hypothesis
    $H_H.append(R_h)$
    $A_S \leftarrow intersection(F_{c_1}, F_{c_2}, F_{c_3})$ Agreed Suggestions
    if $len(A_S) == 20$ **then**
      break
    end
  end
end
return $H_H[-1]$

their assessment, meet these criteria. We only finalize those hypotheses that are agreed upon by all three critics, as we believe that if three state-of-the-art LLMs concur, there is a higher likelihood of the hypothesis being valid. These finalized hypotheses proceed to the evaluation stage. The purpose of this configuration is to test the HGA's ability to generate high-quality hypotheses without receiving any feedback. We aim to maximize both the number of high-quality hypotheses generated (set at 20) and their alignment with the ground truth, by ensuring that only hypotheses agreed upon by all critics are advanced for further evaluation. The method is similar to the approach used in Figure 2 except the knowledge graph and summariser agent is not used.

4.2 Hypotheses Generation with Feedback from Critics

In this configuration (refer Appendix C.2 for the prompt), we again utilize the HGA, three CAs, and introduce a Summarizer Agent (SA). The process begins similarly, with the HGA generating 20 hypotheses based on a goal statement and constraints. The CAs evaluate these hypotheses as before, checking if the generated hypotheses align with the goal and constraints.

However, if not all hypotheses are unanimously agreed upon by the three CAs, each CA generates an individual feedback report for the suggestions it disagrees with, as well as an overall summary. This feedback is then passed to the SA, which synthesizes the reports into a cohesive feedback for the HGA. The HGA uses this feedback to refine its future hypotheses, aiming to reach a consensus among all three CAs.

This feedback loop runs for up to five iterations. If the HGA generates 20 hypotheses that all three CAs agree upon before reaching five iterations, the process is stopped, and the suggestions from that iteration are selected. The goal of this configuration is to maximize the number of high-quality, ground-truth-aligned hypotheses through iterative refinement.The method is similar to the approach used in Figure 2 except the knowledge graph is not used.

4.3 Hypotheses Generation with Knowledge Graph and Feedback from Critics

This configuration (refer to Appendix C.3 for the prompt) builds upon the second setup by incorporating knowledge extracted from a large material science knowledge graph. In addition to the HGA, CAs, and SA, this configuration uses information from the knowledge graph to provide context for the HGA. For this we use MAtKG (Venugopal and Olivetti, 2024) the largest available material science knowledge graph till date. We extract keywords from the goal statement and constraints that relate to specific applications, then use these keywords to identify relevant materials and their properties from the knowledge graph. This extracted information is provided to the HGA as additional context. The idea behind this configuration is to leverage existing knowledge of materials and their properties to help the HGA generate novel and efficient material combinations that achieve the goal while satisfying the constraints. The rest of the process follows the same feedback cycle as described in the second configuration. The Figure 2 depicts this configuration.

5 Novel Evaluation and Metrics

We use a two-part approach to evaluate the generated hypotheses. The first part, closeness, measures how well the hypotheses align with the ground truth

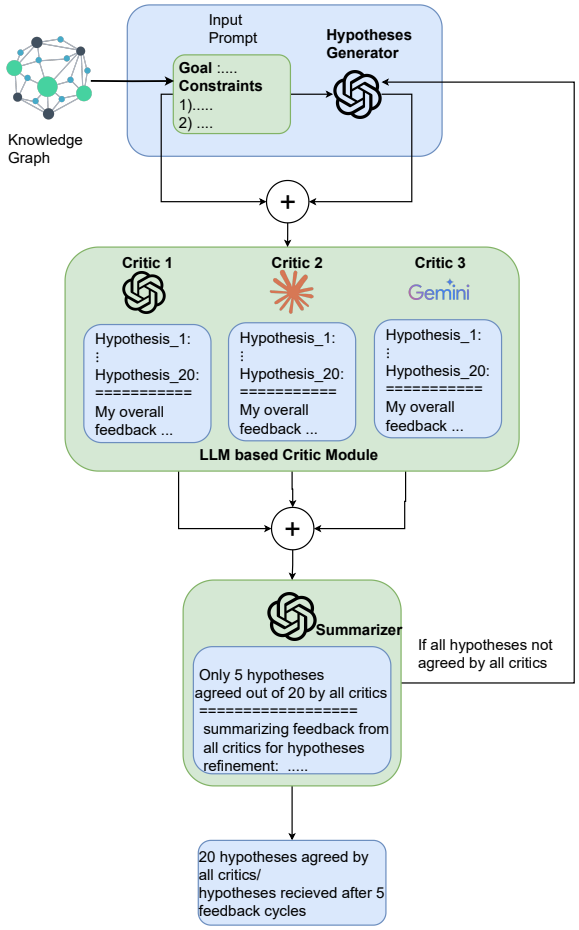

Figure 2: Information from the Knowledge graph, Goal and the constraints are given to the Hypotheses Generator. The generated hypotheses, goal and constraints are provided to the Critic Module for assesment. The Summariser consolidates the feedback from all critics and sends it back to the hypotheses generator if all hypotheses are not agreed else the hypotheses are finalised and sent to the evaluation stage. The feedback loop runs at max for five iterations

in the dataset. The second part, quality, assesses the hypotheses based on six distinct criteria. This dual evaluation is necessary: closeness is applied when a ground truth exists, allowing for direct comparison, while quality is used when only a goal statement and constraints are available, ensuring that hypotheses can still be assessed effectively in the absence of ground truth.

5.0.1 Closeness

The Closeness metric evaluates how similar the generated hypotheses are to the methods in the dataset in terms of concepts, material properties, and key keywords.

1. **Concept Overlap:**
   Evaluates how well the core ideas, methods, and scientific concepts in the generated hypothesis align with those in the ground truth hypothesis.

2. **Property Overlap:**
   Assesses how closely the material properties described in the generated hypothesis align with those in the ground truth, including quantitative values and qualitative descriptions.

3. **Keyword Matching:**
   Evaluates how well specific entities (e.g., material names, chemical compounds, synthesis methods) and keywords in the generated hypothesis match those in the ground truth.

The objective score for evaluating closeness are provided in Table 4

The second part of the evaluation assesses the quality of the generated hypotheses across 6 defined criteria, each specifically tailored to evaluate different aspects of material discovery and design. The detailed description of each criteria and it's scale is mentioned below. The detailed definition associated with the ratings are described in Table 5 (Appendix A).

5.0.2 Quality

This section outlines six key criteria used to measure the quality of the generated hypotheses. The definition of each criterion is provided below:

1. **Alignment with Research Objectives and Constraints:**
   Assesses how directly and effectively the generated hypothesis addresses the objectives specified in the goal statement while adhering to all given constraints and incorporating provided key points or clues.

2. **Scientific Plausibility:**
   Assesses whether the hypothesis is grounded in established scientific principles and theories within material science.

3. **Innovation and Novelty:**
   Measures the degree to which the hypothesis introduces original ideas, perspectives, or methods not previously documented.

4. **Testability:**
   Evaluates how easily and effectively the hypothesis can be tested experimentally, considering the availability of required techniques, equipment, and resources in materials science.

Evaluation Metric	Score 1	Score 2	Score 3	Score 4	Score 5
Concept Overlap	No Overlap: The generated hypothesis contains entirely different concepts.	Minimal Overlap: A few general ideas may be similar, but most key concepts differ.	Moderate Overlap: Some core ideas overlap, but critical concepts are missing or misrepresented.	High Overlap: Most core concepts match with minor differences.	Complete Overlap: The generated hypothesis fully mirrors the core ideas of the ground truth.
Property Overlap	Not Similar: No overlapping or similar properties to the ground truth.	Slightly Similar: A few properties match, but most are different in magnitude or type.	Moderately Similar: Some important properties are similar, though others differ.	Highly Similar: Most key properties align well with the ground truth.	Perfect Match: The generated hypothesis fully matches the properties of the ground truth.
Keyword Matching	No Match: None of the keywords or entities match.	Minimal Match: A small number of non-critical keywords match.	Partial Match: Several important keywords and entities match, but some are missing.	High Match: Most critical keywords and entities match, with minor discrepancies.	Complete Match: All key keywords and entities match exactly with the ground truth.

Table 4: Evaluation Metrics Scores for Closeness between Generated Hypotheses and Ground Truth

5. **Feasibility and Scalability:**
   Evaluates the practicality of implementing the hypothesis at various scales, from laboratory to industrial applications, focusing on existing infrastructure, technology, cost, and effort required to scale the process.

6. **Impact Potential:**
   Assesses the hypothesis's potential to significantly advance the field or address critical challenges in material science.

This evaluation can be weighted, allowing users to customize it based on their specific needs. However, in our approach, we do not apply any weighting to the calculations.

To make evaluation scalable, we use OpenAI o1 as our Evaluation Agent. Given that LLMs have been trained on vast amounts of data, including scientific literature, these models possess an understanding of concepts, material properties, and methods relevant to materials design and discovery to some extent. OpenAI o1, being state-of-the-art for complex reasoning tasks, was a better choice compared to traditional metrics like ROUGE, which struggles to capture semantic similarity. Embedding models, while useful, require fine-tuning on specific datasets and may fail to capture the deep contextual meaning essential for evaluating material science content. To measure closeness, o1 is provided with a prompt that includes the generated hypotheses for a given goal and constraints, along with the corresponding methods from the dataset for evaluation, and the criteria for measuring closeness. The specific prompt used for evaluating closeness can be found in the appendix. The quality is measured in a similar manner, but without the methods from the dataset. The prompt for evaluating quality can be found in the appendix.

6 Results and Analysis

Hypotheses Generation Without Feedback: Lack of Consensus, Less Adherence to Constraints When the Hypotheses Generator is prompted to generate hypotheses without feedback from critics, there is a clear trend where the generated hypotheses often fail to gain unanimous agreement from all critics. Additionally, the generator struggles to adhere to the provided constraints, frequently missing critical details in its reasoning for material selection. The bias towards certain materials and methods results in a narrow scope of exploration, limiting the discovery of alternative solutions.While there were some novel elements in the proposed mechanisms, the hypotheses overall lacked the robustness needed to ensure practical implementation. In terms of methodology, the LLM provided general suggestions which were not accompanied by innovative strategies or detailed procedural guidance. Overall, the generated hypotheses lacked the depth and specificity required for effective experimental application and required further refinement to be practical. This method generates hypotheses which score realtively low on the closeness metric compared to other two techniques. It achieves an average closeness score of 0.7 and an average quality score of 0.8.

Hypotheses Generation with feedback: Improved Reasoning and Diversity of Hypotheses Introducing feedback from critics significantly enhances the quality and diversity of generated hypotheses compared to the method without feedback. With more feedback cycles, the generator better adheres to constraints, producing suggestions more aligned with the goal statements. This approach allows the system to propose a broader range of feasible materials and methods, demonstrating a stronger conceptual understanding. The methodology behind each hypothesis becomes more refined, incorporating feedback to improve reasoning and better meet design goals. Hypotheses generated through this method show a 3% improvement with 0.73 in closeness and a 6.6% increase in quality scores over the previous technique with a score of 0.86. However, despite these improvements, the generated hypotheses often align with well-established methodologies, suggesting familiar materials and techniques. While effective, the reasoning behind material selection lacks depth, focusing on fulfilling design requirements rather than offering innovative or unconventional solutions.

Hypotheses generation with Knowldedge graph and feedback from critics: Novel and feasible This method scores highest on the closeness metric, and generates highest number of hypotheses agreed by all critics on average. With the help of Knowledge graph it is able to provide more diverse combinations of materials which have the potential of pusruing a novel endeavour in synthesizing the suggested materials. Majority of the hypotheses have the potential to be feasible. However in some cases the information from the knoeldege graph which realtes to the goal and constraints provides very fundamental materials which makes the suggestion very broad in a sense and miss the specificity. This method scores 0.8 on closeness, a 7% increase over previous technique and 0.89 on quality scores with 3% increase over previous technique's quality score

6.1 Human Expert Evaluation

A total of 42 suggestions were evaluated by a group of four PhD students in Materials Science to ensure unbiased assessments. These suggestions spanned all three techniques discussed in our paper, with human experts using our evaluation metric to align their judgments with our automated evaluation.

Hypotheses generation without feedback generated hypotheses based on well-established methods and materials, relying heavily on documented techniques. While scientifically sound, experts noted a lack of innovation and weak reasoning for material selection. The hypotheses mainly involved common materials, with limited attention to their interactions, leading to low novelty. Hypotheses generation with feedback showed incremental improvement by incorporating feedback, resulting in more refined suggestions. However, it still focused on well-documented materials, and the reasoning remained shallow. Experts agreed that the combinations were valid but lacked depth and novelty. Hypotheses generation with Knowledge Graph and Feedback received the most positive feedback, offering a wider range of hypotheses and introducing novel material combinations. While some suggestions were too general or lacked domain-specific focus, experts appreciated the creativity and potential for innovation. Despite occasional randomness, this technique showed the most promise in exploring new directions for material design. Overall, the human evaluations aligned well with our automated evaluation, with Hypotheses generation without feedback performing the worst, followed by Hypotheses generation with feedback, and Hypotheses generation with Knowledge Graph and Feedback receiving the highest assessment. This confirms the trends observed in our automated evaluation with evaluation Agent.

7 Conclusion

Our curated dataset plays a crucial role in evaluating LLMs for materials discovery and design, serving as a unique benchmark to assess whether these models generate novel hypotheses or rely on pretraining knowledge. The dataset, composed of 50 papers published after December 2023, ensures the originality of the evaluation. Our LLM-based agentic framework successfully generates novel and feasible hypotheses, though often lacking in-depth reasoning and detailed methodologies needed for implementation. Despite these limitations, the system provides valuable starting points for researchers. The hypotheses generated are scientifically valid but primarily involve well-documented materials and techniques, with some modifications. While effective, the lack of innovation highlights areas for future improvement. Lastly, our evaluation metric, though not at human-level performance, shows strong alignment with human expert assessments, reinforcing its reliability for hypothesis evaluation.

Ethics Statement

We use AI assistants, specifically Grammarly and ChatGPT, were utilized to correct grammatical errors and restructure sentences.

Limitation

The curated dataset, while crucial for ensuring the novelty of generated hypotheses, is limited in size (50 papers) and may not fully capture the diversity of materials science research. Expanding the dataset could lead to more varied and innovative hypotheses, allowing for a more thorough evaluation of the system's capabilities. Additionally, relying solely on LLMs for feedback or critique, without robust checks for hallucinations, poses risks. Even when all critics agree, the suggestions could be scientifically flawed. Our LLM-based agents rely on human-provided constraints to guide their search, but further exploration into how LLMs can autonomously identify relevant constraints is necessary.

References

Chi Chen and Shyue Ping Ong. 2022. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728.

Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 2020. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535.

Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. 2016. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627.

Victor Fung, Shuyi Jia, Jiaxin Zhang, Sirui Bi, Junqi Yin, and Panchapakesan Ganesh. 2022. Atomic structure generation from reconstructing structural fingerprints. *Machine Learning: Science and Technology*, 3(4):045018.

Alireza Ghafarollahi and Markus J Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Geoffroy Hautier, Anubhav Jain, and Shyue Ping Ong. 2012. From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science*, 47:7317–7340.

Ryan Jacobs, Maciej Polak, Lane Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. 2024. Regression with large language models for materials and molecular property prediction.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen T. Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin Aslaug Persson. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1:011002–011002.

Shuyi Jia, Chao Zhang, and Victor Fung. 2024. Llmatdesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*.

Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. 2017. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177.

Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Gutfleisch, and Hongbin Zhang. 2021. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):66.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. 2019. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348.

Rıza Özçelik, Sarah de Ruiter, Emanuele Criscuolo, and Francesca Grisoni. 2024. Chemical language modeling with structured state space sequence models. *Nature Communications*, 15(1):6176.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. 2022. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335.

Sai Krishna Sirumalla, David Stephen Farina Jr, Zhuoran Qiao, Daniele Alessandro Di Cesare, Felipe Costas Farias, Michael Bernard O'Connor, Peter John Bygrave, Feizhi Ding, Thomas Dresselhaus, Marcelo Gomes Pereira de Lacerda, et al. 2024. Multi-modal and multi-task transformer for small molecule drug discovery. In *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*.

Natalia CM Spera, Cristina Salazar-Castro, Paula C Álvarez de Eulate, Yury V Kolenko, and Juliana PS Sousa. 2024. Self–healing core–shell nanofibers for corrosion protective coatings for offshore structures. *Progress in Organic Coatings*, 191:108424.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*.

Vineeth Venugopal and Elsa Olivetti. 2024. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

Mohd Zaki, NM Anoop Krishnan, et al. 2024. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327.

Appendix

A Table of the six quality assement criteria

Evaluation Metric	1	2	3	4	5
Alignment with Research Objectives and Constraints	Misaligned: Hypothesis does not relate to objectives and violates constraints.	Slightly Aligned: Hypothesis touches on the topic but misses key objectives or violates some constraints.	Moderately Aligned: Addresses some objectives but lacks focus on critical elements; meets most constraints.	Highly Aligned: Addresses most objectives and complies with major constraints.	Fully Aligned: Directly addresses all key objectives and fully adheres to all constraints.
Scientific Plausibility	Not Plausible: Contradicts fundamental scientific principles.	Slightly Plausible: Contains significant inaccuracies or unsupported assumptions.	Moderately Plausible: Scientifically sound but includes minor inaccuracies or speculative elements.	Highly Plausible: Scientifically accurate with negligible issues.	Completely Plausible: Fully consistent with established knowledge and well-supported by theory.
Innovation and Novelty	Not Innovative: Restates existing knowledge without introducing new ideas.	Slightly Innovative: Provides minimal new insights or slight variations on known concepts.	Moderately Innovative: Offers some original ideas or novel combinations of existing concepts.	Highly Innovative: Introduces significant new ideas or approaches.	Exceptionally Innovative: Groundbreaking ideas or methodologies that could revolutionize the field.
Testability	Not Testable: Cannot be tested with current techniques.	Difficult to Test: Requires specialized or prohibitively expensive equipment.	Moderately Testable: Can be tested with available techniques but requires complex procedures.	Easily Testable: Testable using common equipment and straightforward procedures.	Highly Testable: Rapid, cost-effective validation with readily available resources and techniques.
Feasibility and Scalability	Not Feasible: Cannot be implemented with current technology or resources.	Slightly Feasible: Theoretically possible but requires breakthroughs or high investment.	Moderately Feasible: Can be implemented on a small scale, but larger-scale production faces challenges.	Highly Feasible: Practical for small and moderate scales with manageable adjustments for scale-up.	Completely Feasible: Easily implemented and scaled from lab to industry with minimal changes or investment.
Impact Potential	No Impact: Unlikely to contribute meaningfully to the field.	Low Impact: Offers minor contributions but lacks significant influence.	Moderate Impact: Can contribute positively in specific areas.	High Impact: Likely to drive significant advancements or address important challenges.	Transformative Impact: Capable of revolutionizing the field or providing solutions to major problems.

Table 5: Evaluation metrics and scales for assessing generated hypotheses based on alignment with research objectives, scientific plausibility, innovation, testability, feasibility, and impact potential.

B Input Prompts

B.1 Goals and Constraints

Goal and constraints provided as input

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. The coating should autonomously recover its protective functionality after mechanical damage and initiate a self-healing process triggered by environmental factors. The method should address challenges related to the implementation of self-healing agents, while providing robust corrosion resistance and allowing for multiple healing cycles.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water) to facilitate immediate and efficient damage recovery in corrosion-prone environments.
2) The self-healing material should allow multiple healing events.
3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage, preventing corrosion from initiating at the damaged site.
4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system, reducing complexity in formulation and ensuring a uniform healing response.

B.2 Prompt for experts list finder

Prompt for experts list finder

Generate a list of experts required to achieve the below mentioned goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. The coating should autonomously recover its protective functionality after mechanical damage and initiate a self-healing process triggered by environmental factors. The method should address challenges related to the implementation of self-healing agents, while providing robust corrosion resistance and allowing for multiple healing cycles.

Just list the top 5 experts in the format "Expert_1, Expert_2, Expert_3, Expert_4, Expert_5"

B.3 Prompt for Hypotheses Generator

Prompt for Hypotheses Generator

Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.
Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective without introducing unsupported information. The ideas should have real-world potential and meet all constraints provided. Provide reason for each suggestion. The suggestions must be in the below mentioned format in a JSON object. For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning: "Give a detailed reasoning for the suggestion"
},
...,
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

B.4 Prompt for Feedback

Prompt for Feedback

Goal_statement:-
{goal_statement}

Constraints:-
{constraint_list}

Suggestions:
generated_suggestions.

Given the goal statement, constraints, and suggestions for materials design and discovery, evaluate each suggestion carefully. Provide detailed, constructive feedback on whether each suggestion meets the goal and satisfies the constraints strictly.

If any suggestion falls short, explain precisely where it fails and suggest specific improvements or adjustments that will help guide the generation agent to produce more accurate and innovative suggestions in the next cycle.

Ensure the feedback is actionable and focused on refining the generation process to align more closely with the goal and constraints in future iterations.

The detailed feedback should be in the below JSON format strictly:

{
"Feedback_for_suggestion_1":
{
"Meets_the_goal_statement_and_satisfies_all_constraints_strictly": "YES/NO",
"Reasoning": " "
},...,
"Feedback_for_suggestion_20":
{
"Meets_the_goal_statement_and_satisfies_all_constraints_strictly": "YES/NO",
"Reasoning": " "
},
"Overall_Feedback_for_improvement_for_future_suggestion_generation": (list of points in a single line)

}

B.5 Prompt for Summarizer

Prompt for Summarizer

Your task is to summarize feedback provided by multiple critics for a list of generated hypotheses. The summary must capture essential points and detailed insights of the feedback generated by all three critics. The final summary generated must be such that it will help in refining future generation of hypotheses such that all the crtics will agree to the newly generated refined suggestions. The feedback is as follows:

Critic_1 feedback:
{feedback_from_claude}

Critic_2 feedback:
{feedback_from_gemini}

Critic_3 feedback:
{feedback_from_gpt}

C Implemented Techniques

C.1 Hypotheses Generation Without feedback

System prompt for Hypotheses Generator

You are an innovative Materials Scientist, Corrosion Engineer, Polymer Chemist, Environmental Scientist, Chemical Engineer capable of doing impactful materials discovery and design.

The expert names are generated by the prompting Hypotheses generator the expert list finder prompt, which it uses as system prompt for itself

User input prompt to the Hypotheses Generator

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. The coating should autonomously recover its protective functionality after mechanical damage and initiate a self-healing process triggered by environmental factors. The method should address challenges related to the implementation of self-healing agents, while providing robust corrosion resistance and allowing for multiple healing cycles.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water) to facilitate immediate and efficient damage recovery in corrosion-prone environments.
2) The self-healing material should allow multiple healing events.
3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage, preventing corrosion from initiating at the damaged site.
4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system, reducing complexity in formulation and ensuring a uniform healing response.
Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.
Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective without introducing unsupported information. The ideas should have real-world potential and meet all constraints provided. Provide reason for each suggestion. The suggestions must be in the below mentioned format in a JSON object. For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning: "Give a detailed reasoning for the suggestion"
},
...,
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

20 Hypotheses Generated by Hypotheses Generator

Suggestion 1:
Materials:Polyurea-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurea matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurea provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.
.
.
.

Suggestion 20:
Materials:Hybrid polymer matrix incorporating both hydrophilic and hydrophobic segments, with embedded nanocapsules containing a water-reactive monomer like methyl methacrylate (MMA).
Methods:Disperse nanocapsules in the hybrid polymer matrix using mechanical mixing. Apply the coating using spray-painting techniques.
Reasoning:MMA polymerizes upon contact with water, providing a quick and effective healing mechanism. The hybrid polymer matrix ensures compatibility with spray-painting and maintains the coating's mechanical integrity.

751

Critic 1 GPT-4o feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurea-based elastomer with embedded microtubules filled with cyanoacrylate offers a scalable and robust self-healing mechanism triggered by water. The co-extrusion process ensures a continuous supply of the healing agent, and the application is compatible with spray-painting. .
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:NO.
Reasoning: Hybrid polymer matrices and nanocapsules introduce formulation complexity. Ensuring uniform nanocapsule distribution and multiple healing events in a scalable manner is challenging.

Overall Feedback for future suggestion improvement:
Focus on single-component healing agents activated by water.
Ensure multiple healing events through vascular-like structures or microtubules
Avoid complex multi-step synthesis processes.
Ensure compatibility with scalable application techniques like spray-painting
Address uniform distribution of self-healing agents within the polymer matrix.

752

Critic 2 Claude 3.5 Sonnet feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: This suggestion meets all constraints. It uses water as a trigger, allows for multiple healing events through microtubules, maintains structural integrity, is compatible with spray-painting, and uses a single-component healing system.
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:NO.
Reasoning: This suggestion fails to meet constraint 2 as it does not provide a mechanism for multiple healing events or a vascular-like structure for continuous supply of the healing agent.

Overall Feedback for future suggestion improvement:
Focus on developing vascular-like structures or continuous supply mechanisms to allow for multiple healing events
Ensure all suggestions explicitly address each constraint, particularly the requirement for multiple healing cycles
Explore innovative ways to integrate vascular structures with spray-painting techniques for large-scale application
Consider combining successful elements from different suggestions, such as microtubules or vascular networks with various healing agents
Investigate methods to create self-replenishing systems within the coating that are compatible with spray application

Critic 3 Gemini 1.5 flash feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:The suggestion effectively addresses the goal and constraints. The use of polyurea for durability, cyanoacrylate for rapid healing, and PLA microtubules for continuous healing agent supply all contribute to a robust self-healing system. The co-extrusion method ensures scalability, and the spray-painting application aligns with industrial requirements.
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: The polyester coatings with embedded microcapsules containing water-activated DMPA offer good mechanical properties and chemical resistance for offshore applications. The spray-painting compatibility and the single-component healing agent make this suggestion a strong candidate.

Overall Feedback for future suggestion improvement:
Explore a broader range of materials, consider incorporating bio-inspired self-healing mechanisms, and investigate the use of advanced characterization techniques to validate the proposed solutions.

Finalised Suggestions Agreed by all Critics

Suggestion 1:
Materials:Polyurea-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurea matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurea provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

Suggestion 17:
Materials:Polyurethane-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurethane matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurethane provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

Suggestion 9:
Materials:Polyurethane-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurethane matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurethane provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

C.2 Hypotheses Generation With feedback from Critics

Summarised feedback from Summariser fed to the Hypotheses Generator for refinement

Summary of Feedback from Multiple Critics on Generated Hypotheses:
Common Approvals and Rejections: Approval Trends: Polyurethane matrix with cyanoacrylate and epoxy resin with TMPTA were universally approved for their scalability, effectiveness in multiple healing events, and compatibility with industrial applications like spray-painting. Fluoropolymer coatings with n-butyl acrylate-filled channels** and **polyurea-based elastomer with polyisocyanate-filled channels were consistently approved for their chemical resistance and suitability for harsh environments, as well as their scalable application methods.
.
.
.
Single-component healing systems triggered by water were highly favored across all suggestions. For future hypothesis generation, focus on combining simple, scalable manufacturing processes with innovative material choices and healing mechanisms. Prioritize single-component, water-activated healing agents compatible with industrial application techniques. Incorporate advanced characterization methods and additional functionalities to refine and validate the solutions robustly. Exploring hybrid systems and alternative stimuli can further enhance the effectiveness and adaptability of self-healing coatings.

Refined Hypotheses Generated by Hypotheses Generator by incoprorating feedback

Suggestion 1:
Materials:Polyurethane matrix with cyanoacrylate-filled vascular-like channels.
Methods:3D print a network of channels within the polyurethane matrix using a sacrificial material like PVA. Dissolve the PVA post-formation to leave hollow channels and inject cyanoacrylate. Apply the final coating using spray-painting techniques.
Reasoning:Polyurethane provides durability while cyanoacrylate offers a rapid water-activated healing mechanism. The vascular-like structure allows multiple healing events, and the method is compatible with scalable techniques like spray-painting.
.
.
.
Suggestion 20:
Materials:Polyurethane matrix with embedded microcapsules containing isocyanate.
Methods:Disperse microcapsules containing isocyanate uniformly in a polyurethane matrix. Apply the coating using spray-painting techniques. Upon damage, isocyanate reacts with water to form a protective polymer.
Reasoning:Isocyanate offers a simple and efficient water-activated healing mechanism. Polyurethane provides durability, and microcapsules ensure even distribution of the healing agent.

Critic 1 GPT-4o feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurethane matrix with cyanoacrylate-filled vascular-like channels is durable, has a rapid water-activated healing mechanism, allows multiple healing events, and is compatible with scalable techniques like spray-painting.. .
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurethane matrix with isocyanate-filled microcapsules ensures a simple, efficient water-activated healing mechanism, durability, and compatibility with spray-painting techniques. .

Overall Feedback for future suggestion improvement:
All of the suggestions satisfy the constraints and align with the goals

Critic 2 Claude 3.5 Sonnet feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:This suggestion meets all constraints. It uses water as a trigger for cyanoacrylate, allows for multiple healing events through vascular-like channels, maintains structural integrity with polyurethane, is compatible with spray-painting, and uses a single-component healing system.
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: This suggestion satisfies all constraints. It uses water as a trigger for isocyanate, provides multiple healing events through microcapsules, maintains structural integrity with polyurethane, is compatible with spray-painting, and uses a single-component healing system..

Overall Feedback for future suggestion improvement:
I agree with all of the suggestion, but they can be further improved :
Explore novel combinations of matrix materials and healing agents to enhance performance and durability .....

Critic 3 Gemini 1.5 flash feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:This suggestion effectively addresses the goal and constraints. The use of polyurethane for durability, cyanoacrylate for rapid healing, and PVA microtubules for continuous healing agent delivery all contribute to a robust self-healing system. The 3D printing method ensures scalability, and the spray-painting application aligns with industrial requirements.
.
.
.
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:Fluoropolymers offer chemical resistance, and n-butyl acrylate provides a water-activated healing mechanism. The method is scalable and effective.

Overall Feedback for future suggestion improvement:
All suggestions align with constrainst and goals, it can be further improved .....

C.3 Hypotheses Generation With Knowledge Graph and Feedback from Critics

User input prompt to the Hypotheses Generator with Knowledge Graphs

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. The coating should autonomously recover its protective functionality after mechanical damage and initiate a self-healing process triggered by environmental factors. The method should address challenges related to the implementation of self-healing agents, while providing robust corrosion resistance and allowing for multiple healing cycles.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water) to facilitate immediate and efficient damage recovery in corrosion-prone environments.
2) The self-healing material should allow multiple healing events.
3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage, preventing corrosion from initiating at the damaged site.
4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system, reducing complexity in formulation and ensuring a uniform healing response.
Relevant information from trusted Knowledge Graphs:

Suggested Materials:
Aluminum: Aluminum is known for its corrosion resistance and is commonly used in offshore environments.
Graphene: Graphene has excellent barrier properties and mechanical strength, making it suitable for self-healing coatings.
Epoxy: Epoxy is widely used in coatings for its strong adhesion and durability, essential for corrosion protection.
Zinc: Zinc is often used in coatings for its sacrificial protection properties, which are beneficial for corrosion resistance.

Suggested Properties:
Corrosion Resistance: Essential for protecting metallic structures in corrosive offshore environments.
Barrier Properties: Important for preventing the ingress of corrosive agents like water and salts.
Mechanical Property: Necessary to ensure the coating can withstand mechanical damage and maintain its integrity.
Adhesions: Critical for ensuring the coating remains attached to the substrate, especially in harsh conditions.

TASK:
Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. Use the information provided from the trusted knowledge graph as a reference to generate suggestions, but aim to propose novel combinations or entirely new materials and methods.While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.

Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective without introducing unsupported information. The ideas should have real-world potential and meet all constraints and the goal provided. Provide scientific reasoning for each suggestion.The suggestions must be in the below mentioned format in a JSON object For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning: "Give a detailed reasoning for the suggestion"
},
...,
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

20 Hypotheses Generated by Hypotheses Generator with Knowledge Graphs

Suggestion 1:
Materials:A single-component polyurethane matrix combined with zinc oxide nanoparticles and tannin-derived microcapsules. The tannin-derived microcapsules offer a simple, natural healing mechanism activated by water exposure, while zinc oxide provides dual protection and healing properties.
Methods:Utilize a scalable spray-painting method to apply the polyurethane coating. The zinc oxide nanoparticles enhance the barrier properties, and the tannin-derived microcapsules release their content upon contact with water, initiating self-healing.
Reasoning:This design leverages simplicity and sustainability by using tannin-derived agents, aligning with critics' feedback to avoid complex silane-based solutions. Zinc oxide offers excellent corrosion resistance, making it suitable for offshore environments.

Suggestion 2:
Materials:Epoxy matrix with embedded graphene oxide and water-responsive natural polymers, such as chitosan. Graphene oxide provides excellent barrier properties, while chitosan acts as a natural, biodegradable healing agent.
Methods:Develop a sprayable epoxy coating where graphene oxide enhances mechanical strength. Chitosan microcapsules release upon water contact, enabling multiple healing cycles.
Reasoning:Graphene oxide enhances both mechanical and barrier properties. The use of chitosan aligns with sustainability goals, offering a biodegradable option that critics favor.
.
.
.
Suggestion 20:
Materials:Polysiloxane matrix with embedded graphene oxide and tannin-derived microcapsules.
Methods:Create sprayable polysiloxane coatings where graphene oxide enhances barrier properties. Tannin-derived microcapsules release healing agents upon water exposure.
Reasoning:This suggestion uses natural healing agents and graphene oxide for enhanced performance, aligning with simplicity and environmental considerations.