From Text Generation to Structured Reasoning: Structure-Aware Hypothesis Evolution for Scientific Hypothesis Generation

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown promise for automated scientific hypothesis generation, yet most existing approaches formulate hypothesis generation at the text level, without explicitly modeling the structured reasoning processes underlying scientific discovery. As a result, hypothesis generation and refinement are often weakly grounded in scientific logic and misaligned with expert reasoning.

We propose a structure-centric framework that models scientific hypotheses as hierarchical reasoning chains. To support structure-aware modeling and evaluation, we introduce HSRC-5000, a large-scale materials science dataset constructed by decomposing scientific papers into explicit reasoning components. Building on this representation, we construct the Hierarchical Scientific Reasoning Graph and propose Structure-Aware Hypothesis Evolution (SAHE), a dimension-aware evolutionary framework that enables controlled and causally consistent hypothesis generation.

Experiments show that explicitly modeling scientific reasoning structure consistently improves hypothesis generation quality, yielding an average absolute improvement of about one point in overall score over strong retrieval- and evaluation-based baselines, with particularly pronounced gains in logical coherence and multi-dimensional novelty. Qualitative analyses further indicate close alignment between the generated reasoning chains and human expert scientific logic. The code and dataset will be publicly released.[1]

1 Introduction

Scientific discovery hinges on posing meaningful questions, formulating plausible hypotheses, and validating them through systematic investigation. Traditionally, transformative hypotheses emerge through the accumulation of domain expertise and specialized cognitive processes. Recent advances in Large Language Models (LLMs) have positioned AI-driven scientific discovery as an emerging research frontier, leveraging LLMs' capabilities in abstraction, reasoning, and generative inference.

LLMs have shown growing promise in scientific idea and hypothesis generation. For example, SI et al. (Si et al., 2025) demonstrated that model-generated research ideas can achieve novelty levels comparable to those proposed by human researchers. Related studies have explored LLM-driven hypothesis generation across diverse domains, including the social sciences (Zhou et al., 2024), nanobody design (Swanson et al., 2025), algorithm optimization (Novikov et al., 2025), and automated research workflows (Weng et al., 2025).

Despite recent progress, current approaches to scientific hypothesis generation exhibit fundamental limitations. Most methods treat hypothesis generation as a conventional text generation problem, even when combined with retrieval or evolutionary refinement, operating primarily over surface-level natural language rather than explicitly modeling the underlying scientific reasoning relationships. Such approaches are particularly inadequate in domains such as chemistry and materials science, where valid hypotheses depend on coherent causal chains, theoretical justification, and experimental constraints. In scientific practice, hypotheses arise from interconnected reasoning processes that link problem formulation, methodological choices, and expected outcomes.

However, existing methods largely ignore these reasoning relationships, relying instead on coarse-grained text-level operations such as mutation and crossover, which can easily disrupt causal consistency and lead to logically fragile hypotheses. Moreover, most available datasets emphasize titles and abstracts, providing limited access to experi-

[1]The project repository will be available at `https://anonymous.4open.science/r/HSRC/`.

mental design and reasoning context, further constraining models' ability to generate hypotheses grounded in scientific logic.

To address these challenges, we propose a structured, multi-dimensional framework for scientific hypothesis generation. We introduce the *Hierarchical Scientific Reasoning Chain (HSRC)*, which formalizes scientific reasoning into five interdependent components: *Target*, *Strategy*, *Method*, *Result*, and *Heuristic*. Together, these dimensions capture the logical structure underlying scientific discovery, from research objectives to generalizable insights. Based on this framework, we make three primary contributions:

First, we introduce **HSRC-5000**, the first large-scale dataset designed to support structure-aware scientific hypothesis generation in materials science. By decomposing over 5,000 research articles into explicit reasoning components, HSRC-5000 fills a critical gap in existing materials datasets, which typically lack fine-grained representations of experimental reasoning, procedural decisions, and result interpretation.

Second, we construct the ***Hierarchical Scientific Reasoning Graph (HSRC-Graph)***, a heterogeneous knowledge graph that integrates HSRC reasoning chains through shared scientific entities. This representation enables hypothesis generation as structured reasoning over both intra-paper causal relations and inter-paper entity connections.

Third, we propose ***Structure-Aware Hypothesis Evolution (SAHE)***, the first dimension-aware hypothesis generation framework that explicitly operates over individual reasoning dimensions. By applying constrained variation operators within each dimension, SAHE preserves causal consistency while enabling controlled exploration of the hypothesis space, supporting the generation of diverse hypotheses evaluated along validity, novelty, feasibility, and potential impact.

Empirically, extensive experiments on the HSRC-5000 benchmark demonstrate that explicitly modeling scientific reasoning structure yields substantial improvements in overall hypothesis quality, outperforming strong retrieval- and evaluation-based baselines as well as the state-of-the-art chemistry-domain method **MOOSE-CHEM**.

By framing hypothesis generation as a structured, multi-dimensional reasoning problem, our work emphasizes the central role of explicit reasoning structure in AI-driven scientific discovery and provides a foundation for more reliable and interpretable hypothesis generation.

2 Related Work

Recent advances in Large Language Models (LLMs) have enabled a common pipeline for automated scientific hypothesis generation, typically involving external knowledge retrieval, hypothesis generation with iterative refinement, and LLM-based evaluation or ranking (Zhou et al., 2024; Si et al., 2025). While task-specific designs vary, most existing systems operate primarily at the level of natural language text, with limited explicit modeling of the underlying scientific reasoning structure. We review representative approaches below.

2.1 Knowledge Retrieval Mechanisms

Access to relevant domain knowledge is a prerequisite for effective hypothesis generation. Most existing approaches rely on document- or text-level retrieval. For example, MOOSE-Chem (Yang et al., 2025) adopts a sliding-window strategy with multi-round filtering to retrieve chemistry literature, while SI et al. (Si et al., 2025) employ a Retrieval-Augmented Generation (RAG) framework with iterative re-ranking. Systems such as AlphaEvolve (Novikov et al., 2025) and CycleResearcher (Weng et al., 2025) further integrate retrieval to expand candidate solution spaces.

Beyond text retrieval, several works explore structured representations to improve knowledge access. KG-CoI (Xiong et al., 2024) incorporates knowledge-grounded modeling to capture relationships among scientific concepts, while multi-agent systems such as Virtual Lab (Swanson et al., 2025) and physics reasoning agents (Xu et al., 2025) leverage entity-level structures for coordination and information sharing. However, in these approaches, structured representations are primarily used to facilitate retrieval or agent interaction, rather than to explicitly model the internal reasoning components that constitute a scientific hypothesis.

As a result, existing retrieval mechanisms generally lack fine-grained, dimension-aware access to reasoning elements such as research objectives, strategies, methods, and experimental outcomes, which are central to expert scientific reasoning.

2.2 Hypothesis Generation and Iteration

For hypothesis generation and iteration, evolutionary and sampling-based strategies are

widely adopted. MOOSE-Chem and AlphaEvolve (Novikov et al., 2025) employ genetic algorithms or MAP-Elites-style methods, iteratively modifying hypotheses via mutation and crossover. These operators, however, are typically applied directly to unstructured text, without explicit awareness of dependencies among different reasoning components, making it difficult to preserve causal consistency.

Alternative approaches emphasize large-scale sampling and filtering. SI et al. (Si et al., 2025) generate extensive candidate sets followed by semantic deduplication and ranking, while Zhou et al. (Zhou et al., 2024) treat hypothesis generation as a text generation task guided by LLM-based evaluation. Although effective for broad exploration, these methods remain largely quantity-driven and do not explicitly optimize individual reasoning components.

More structured efforts, including knowledge-grounded modeling and chain-based reasoning, attempt to incorporate specific reasoning patterns, and multi-agent systems (Swanson et al., 2025; Xu et al., 2025) further enhance task decomposition. Nevertheless, hypothesis generation in these approaches is still predominantly realized through holistic text production, limiting interpretability and control at the level of scientific reasoning structure.

Overall, existing methods either emphasize exploration over structure or rely on implicit representations of reasoning, highlighting the need for explicit, dimension-aware modeling of scientific reasoning to support more reliable and interpretable hypothesis generation.

3 Dataset Construction

To address the lack of publicly available and fine-grained datasets for scientific hypothesis generation in materials science, we construct HSRC-5000, a full-document-level corpus derived from complete research articles. Existing benchmarks in this domain are either abstract-level or limited to flattened textual representations, which makes it difficult to capture the internal organization of scientific reasoning across a full study.

HSRC-5000 is designed to preserve the structural signals of scientific problem formulation, experimental design, and empirical findings within full papers. While the dataset itself does not assume a specific reasoning formalism, it is organized in a way that enables downstream models to analyze and generate hypotheses with explicit reasoning structure. The Hierarchical Scientific Reasoning Chain (HSRC) serves as a conceptual framework for organizing and interpreting these structures, and will be formally introduced in the next section.

3.1 Data Acquisition and Preprocessing

We collected 10,213 materials science papers published between 1991 and 2025 from authoritative academic databases. To ensure data quality and domain relevance, we applied a multi-stage filtering process:

Quality Control. We retained only papers published in JCR Q1 or Q2 journals, excluding low-quality or potentially predatory venues.

Domain Validation. A large language model–based classifier was used to assess semantic relevance from abstracts, removing papers only marginally related to materials science.

Metadata Normalization. Bibliographic metadata were completed via DOI-based queries to the Crossref API, and author, affiliation, and citation fields were standardized.

After filtering, the final dataset contains 5,128 high-quality papers from 469 journals, with 36.8

3.2 Structured Reasoning Chain Extraction

As scientific articles are distributed primarily in PDF format, we first converted all papers into high-fidelity Markdown using MinerU (Wang et al., 2024), followed by denoising of experimental sections to remove citation markers and non-textual artifacts.

We then extracted structured information corresponding to the five HSRC dimensions—*Target*, *Strategy*, *Method*, *Result*, and *Heuristic*—from the full text under strict definition constraints. To reduce hallucination, we introduced an explicit missing-information mechanism that allows the model to output `Insufficient Information` when a dimension cannot be reliably inferred, rather than forcing speculative completion. As a result, more than 90

3.3 Dataset Splitting and Statistics

To emulate a realistic scientific discovery setting—leveraging historical knowledge to address emerging research problems—we partitioned the dataset strictly along the temporal axis:

- **Retrieval Corpus**: 5,003 papers published

prior to 2023, used as the knowledge base for retrieval and reasoning pattern learning.

- **Test Set**: 125 papers published in or after 2023, designed to evaluate hypothesis generation for frontier research questions.

This time-aware split prevents data leakage and provides a realistic assessment of generalization in scientific frontier exploration. Detailed dataset statistics are reported in Figure 5 in the appendix.

4 Methodology

4.1 Theoretical Foundation: From Cognitive Inquiry to HSRC

Classical cognitive theories model scientific problem-solving as a cyclical process involving problem perception, hypothesis formation, reasoning analysis, and empirical validation. Dewey's reflective thinking emphasizes the dynamic progression from problem to hypothesis to validation, while Polya's four-step framework elaborates the process through problem understanding, plan formulation, execution, and reflection. Both highlight the pivotal role of transferable heuristic knowledge in scientific progress.

Based on these insights, we abstract these steps into a hierarchical scientific reasoning chain (HSRC), capturing the internal logic of scientific inquiry. We formalize a scientific hypothesis H as a tuple defined on a structured logical manifold:

$$H = \langle \mathcal{T}, \mathcal{S}, \mathcal{M}, \mathcal{R}, \mathcal{H} \rangle \quad (1)$$

Each component addresses a distinct scientific question:

- **Target** ($\mathcal{T}$): *What is the research objective?* Corresponding to Polya's "understand the problem," it defines the motivation, performance metrics, and application context.

- **Strategy** ($\mathcal{S}$): *What principles guide the study?* Corresponding to "devise a plan," it represents high-level scientific intuition or design philosophy (e.g., interface engineering, band-gap tuning).

- **Method** ($\mathcal{M}$): *How is the plan executed?* Corresponding to "execute the plan," it includes experimental parameters, synthesis procedures, and characterization techniques.

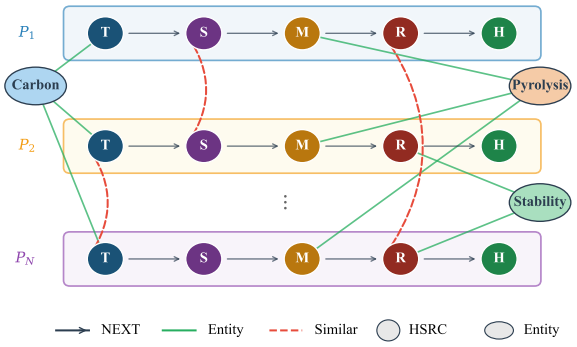

Figure 1: Structure of the Hierarchical Scientific Reasoning Graph (HSRG).

- **Result** ($\mathcal{R}$): *What outcomes are obtained?* Corresponding to Dewey's "empirical validation," it provides quantitative feedback and mechanistic explanations.

- **Heuristic** ($\mathcal{H}$): *What transferable insights emerge?* Corresponding to Polya's "reflection," it captures generalizable heuristics and design principles.

This formalization transforms hypothesis generation from holistic text production into structured reasoning over explicit components, enabling search, comparison, and evolution.

4.2 Hierarchical Scientific Reasoning Graph (HSRG)

Scientific breakthroughs often arise from integrating knowledge across multiple studies. To capture this, we construct a Hierarchical Scientific Reasoning Graph (HSRG) that unifies reasoning chains into a heterogeneous knowledge graph.

4.2.1 Chain-Entity Dual-Layer Architecture

Standard scientific knowledge graphs typically focus on static entity relationships and overlook processual reasoning. HSRG decouples reasoning chains from semantic entities:

Reasoning Chain Layer: Each paper corresponds to a directed HSRC chain $\mathcal{T} \to \mathcal{S} \to \mathcal{M} \to \mathcal{R} \to \mathcal{H}$, capturing causal and temporal dependencies.

Semantic Entity Layer: Entities such as materials, methods, and heuristics are shared across papers and anchored to reasoning stages, enabling fine-grained alignment.

Figure 1 illustrates the structure of the HSRG.

4.2.2 Graph Construction

Entity Grounding. A few-shot LLM-based extractor identifies domain-specific entities from HSRC text segments. Extracted entities are unified via dictionary-based normalization and string similarity clustering to reduce lexical variance.

Reasoning Chain Alignment. Each HSRC node is encoded using MatSciBERT. Semantic similarity edges are induced by cosine similarity between nodes of the same reasoning dimension, ensuring that alignment is performed among semantically comparable reasoning components and resulting in a sparse yet coherent graph.

The HSRG provides a unified knowledge space for structured hypothesis evolution.

4.3 Structure-Aware Hypothesis Evolution (SAHE)

We model scientific hypothesis generation as a constrained multi-objective optimization problem over the HSRC-induced space $\mathcal{S}_{\text{HSRC}}$. A hypothesis is represented as a structured reasoning tuple

$$\mathcal{H} = \langle T, S, M, R, H \rangle, \tag{2}$$

where H encodes transferable heuristic knowledge that modulates the evolution of the remaining reasoning components.

The optimization objective is defined over the instantiated reasoning chain $\mathcal{I} = \{T \rightarrow S \rightarrow M \rightarrow R\}$, subject to heuristic consistency:

$$\begin{aligned} &\max_{\mathcal{I} \in \mathcal{S}_{\text{HSRC}}} \quad \mathbf{f}(\mathcal{I}) = \big(\phi_{\text{logic}}, \phi_{\text{novelty}}, \phi_{\text{feasibility}}, \phi_{\text{impact}}\big), \\ &\text{s.t.} \quad \mathcal{C}(\mathcal{I}, H) = \text{true}. \end{aligned} \tag{3}$$

4.3.1 Initialization

Given a research question q, we retrieve K highly relevant seed papers from the HSRG. Each paper is parsed into its HSRC representation, which serves as a structured reasoning template. The LLM then generates multiple candidate hypotheses, and a logical consistency scoring function filters out incoherent candidates, yielding the initial hypothesis population.

4.3.2 Structure-Aware Evolutionary Operators

Unlike traditional evolutionary methods that treat hypotheses as unstructured text, SAHE operates on the structured HSRC representation, applying operators to individual reasoning dimensions while maintaining causal consistency.

Target Mutation Refines or expands the research objective based on information from the T-path, ensuring downstream stability by keeping (S, M, R) unchanged.

Strategy Transplant Introduces alternative strategies from the S-path, using a compatibility function to verify feasibility under the current target.

Method Recombination Explores the method subspace using papers from the M-path, allowing either full replacement or partial fusion of experimental procedures.

Material-Specific Inspiration For hypotheses involving the same material, this operator incorporates alternative processing or experimental perspectives from related studies, without altering the material itself.

4.3.3 Dimension-Constrained Crossover and Heuristic Fusion

Dimension-Constrained Crossover: To avoid violating causal dependencies, crossover between parent hypotheses $\mathcal{I}_a$ and $\mathcal{I}_b$ is strictly limited to whole-dimension exchanges (e.g., $S_a \leftrightarrow S_b$).

Heuristic-Guided Fusion: Beyond simple exchange, SAHE leverages the heuristic dimension $\mathcal{H}$ as a *high-order control signal*. Rather than being rewritten independently, $\mathcal{H}$ constrains the joint evolution of T, S, M, R. This mechanism allows cross-dimensional evolution to be modulated by transferable scientific paradigms, significantly improving creativity while maintaining global causal consistency.

4.3.4 Evaluation and Selection

In each generation, an LLM evaluates hypotheses on the multi-objective criteria defined above and provides refinement suggestions. We employ an elitist strategy (retaining top 10% Pareto-optimal candidates) and semantic deduplication to maintain diversity. The final set is produced via tournament selection after T iterations.

5 Experiments

5.1 Experimental Setup

We systematically evaluate the proposed method on the automated scientific hypothesis generation task in materials science. Given a research question q, the model is tasked with generating hypotheses that

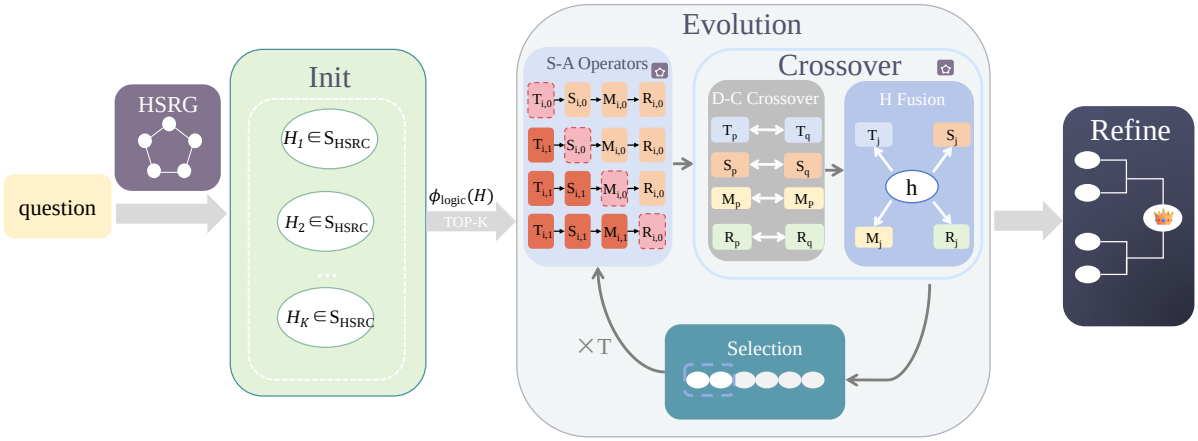

Figure 2: Overview of the Structure-Aware Hypothesis Evolution (SAHE) framework. Hypotheses are initialized from structure-aware retrieval over HSRG, evolved through dimension-specific operators and constrained crossover, and iteratively refined via evaluation-driven selection.

demonstrate *novelty, logical consistency, feasibility*, and *potential scientific impact*.

Dataset. **HSRC-5000 Test Set**.

Model and Implementation Details. All methods use the same backbone LLM (GPT-5.1) to ensure fair comparison. For retrieval-based baselines, the literature corpus, encoder, and number of retrieved documents K are fixed. For evolutionary methods, population size and iteration number are identical. Scores are aggregated at the question level; reported mean and standard deviation are computed across all test questions.

Baselines and Variants. We consider the following baselines and ablation variants:

- **Zero-shot**: Direct LLM generation without retrieval or evaluation.
- **RAG-only**: Vector-based retrieval augmentation without structured reasoning or evolution.
- **HSRG + Eval (No SAHE)**: Structured hypothesis representation with iterative evaluation–refinement cycles matched in number to SAHE iterations, but without population-based evolutionary operators.
- **HSRG + SAHE, w/o H**: Full evolutionary framework without the heuristic dimension.
- **MOOSE-CHEM** (ICLR 2025): A chemical-domain hypothesis generation method.
- **HSRG + SAHE**: Full method with hierarchical representation and heuristic-guided evolution.

Evaluation Metrics. Generated hypotheses are evaluated on logical consistency (**Logic**), novelty across dimensions (**Nov-T/S/M/R**), experimental and engineering feasibility (**Feasibility**), and potential scientific impact (**Impact**). Metrics are scored by an evaluation model under standardized criteria. It is important to note that for a given research question q, multiple valid solutions may exist. Consequently, hypothesis quality cannot be determined solely by comparison with future publications or a single reference outcome. Our multi-dimensional evaluation considers both structural reasoning alignment and internal consistency, ensuring that creative yet valid hypotheses are not misjudged as incorrect.

5.2 Main Results and Ablation Analysis

Table 1 summarizes the main and ablation experiment results on the HSRC-5000 Test Set.

Overall Performance. Our full method (**HSRG + SAHE**) achieves the best or tied-best scores across all metrics, particularly excelling in logical consistency, novelty, and overall performance. This confirms that structured collaboration between retrieval, evaluation, and evolutionary modules is key to high-quality hypothesis generation.

Comparison with MOOSE-CHEM. We compare our method with MOOSE-CHEM, a recent hypothesis generation approach originally proposed

for the chemical domain. While chemistry and materials science share similar scientific paradigms, MOOSE-CHEM treats hypothesis generation primarily as a text-level generation problem. In contrast, our method performs dimension-aware hypothesis generation, explicitly modeling and evolving hypotheses along multiple scientific dimensions. MOOSE-CHEM achieves competitive results, but our method consistently performs better in logical consistency, novelty, and overall score, highlighting the advantage of structured, dimension-aware reasoning.

Impact of Retrieval and Structured Representation. Comparison between **Zero-shot** and **RAG-only** indicates that retrieval alone does not consistently improve performance. Without structured modeling and controllable reorganization, additional context may not be effectively used and can even interfere with reasoning. This highlights the limitations of naive RAG in complex scientific reasoning.

Evaluation Mechanism versus Evolution. **HSRG + Eval (No SAHE)** shows improved performance relative to pure retrieval, as it performs repeated evaluation-and-improvement iterations equivalent to SAHE, but lacks evolutionary operators. The performance gap between this baseline and the full model demonstrates that evolutionary search provides global exploration and diversity beyond local evaluation corrections.

Role of the Heuristic Dimension. Removing the heuristic dimension (**HSRG + SAHE, w/o H**) reduces novelty and overall score, while logical consistency remains largely unchanged. This confirms that the heuristic dimension functions as a high-order control signal guiding creative exploration rather than directly ensuring causal correctness.

Summary. Scientific hypothesis generation is fundamentally a structure-constrained search problem, not mere text generation. Structured scientific representation offers actionable reasoning units, evaluation provides directional feedback, and heuristic-guided evolutionary search enables efficient exploration. The synergy of these components balances novelty, logical consistency, and feasibility.

5.3 Evolutionary Dynamics Analysis

We recorded population dynamics across generations to validate convergence and diversity. Figure 3 shows steady improvement across all metrics, with the largest gains in the final generation, demonstrating the effectiveness of crossover and mutation in integrating advantageous features.

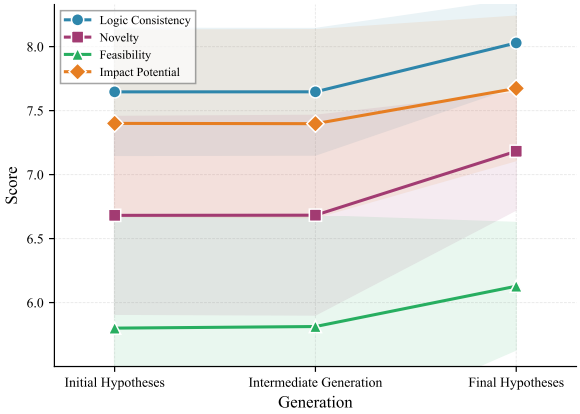

Figure 3: Evolution of evaluation metrics across generations.

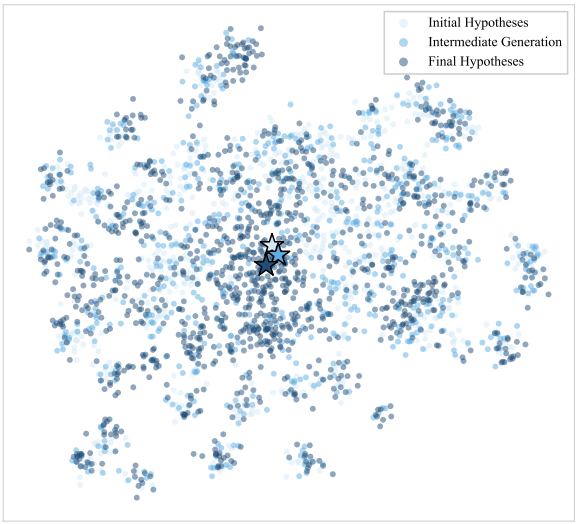

Figure 4: t-SNE visualization of hypothesis population, with stars indicating centroids.

Figure 4 visualizes the population in semantic space via t-SNE. Final hypotheses (deep blue) maintain broad coverage, while centroids (stars) move progressively. This indicates a balance between exploration and exploitation, preserving diversity while improving scores.

5.4 Structural Alignment with Human Expert Reasoning

Case Analysis. This example demonstrates that the generated hypothesis aligns with human expert reasoning not only superficially but structurally across the HSRC dimensions.

- **Target dimension:** Both the human study

Table 1: HSRC-5000 Test Set: Performance Comparison of Hypothesis Generation Methods

Method	Logic	Nov-T	Nov-S	Nov-M	Nov-R	Feasibility	Impact	Overall
Zero-shot	7.16 ± 1.15	6.12 ± 1.35	6.62 ± 1.19	5.89 ± 1.17	6.88 ± 1.20	5.58 ± 1.17	7.25 ± 1.25	6.38 ± 1.09
RAG-only	6.77 ± 2.87	5.89 ± 2.56	6.50 ± 2.79	5.95 ± 2.58	6.36 ± 2.73	5.08 ± 2.22	6.58 ± 2.82	6.17 ± 2.63
HSRG + Eval (No SAHE)	5.48 ± 4.08	4.46 ± 3.37	5.57 ± 4.16	5.13 ± 3.84	4.93 ± 3.68	4.22 ± 3.16	5.13 ± 3.82	5.02 ± 3.74
MOOSE-CHEM (ICLR 2025)	6.85 ± 1.18	6.25 ± 1.28	6.45 ± 1.35	6.20 ± 1.38	6.30 ± 1.25	5.45 ± 1.10	6.65 ± 1.20	6.40 ± 1.18
HSRG + SAHE, w/o H	7.68 ± 1.24	6.40 ± 1.33	7.29 ± 1.36	6.57 ± 1.32	6.91 ± 1.32	5.66 ± 1.06	7.10 ± 1.36	6.79 ± 1.25
HSRG + SAHE	**8.10 ± 0.32**	**7.07 ± 0.67**	**7.87 ± 0.44**	**7.13 ± 0.60**	**7.67 ± 0.49**	**6.34 ± 0.54**	**7.90 ± 0.46**	**7.43 ± 0.37**

Table 2: Alignment between human expert reasoning and model-generated hypothesis under the HSRC framework for cryogenic SPDT of PMMA.

Dimension	Human Expert (Published Study)	Ours (Generated Hypothesis)
Target	Improve surface form accuracy and surface roughness of PMMA in ultra-precision SPDT for optical components	Improve surface form accuracy and surface roughness of PMMA in ultra-precision SPDT, extended to high-precision polymer components with explicit consideration of process robustness
Strategy	Apply cryogenic cooling to suppress viscoelastic effects and enhance material rigidity during cutting	Apply controlled cooling to suppress viscoelastic relaxation and stabilize cutting, enabling systematic exploration of temperature–mechanics interactions
Method	Perform cryogenic SPDT at 0 °C; optimize cutting parameters via Taguchi design; characterize temperature-dependent mechanics using nanoindentation and DMA	Preserve the underlying experimental rationale revealed by mechanical characterization, but reformulate it as a descriptor-centered design that normalizes material response across temperatures and cutting conditions
Result	Cryogenic cutting yields lower surface roughness and form error than room temperature machining, with increased hardness and modulus	Cryogenic conditions consistently improve surface quality and form accuracy, with results organized into generalizable machinability regimes beyond specific parameter settings

and the model focus on improving PMMA surface quality, with the model extending the scope to additional polymer applications and robustness considerations.

- **Strategy dimension:** The core idea of cryogenic cooling to suppress viscoelastic effects is preserved, with the model additionally enabling systematic exploration of parameter interactions, showing a deeper abstraction of scientific strategy.

- **Method dimension:** While the human study specifies concrete experimental parameters, the model abstracts these into a descriptor-centered design that captures the mechanical response across varying conditions, maintaining the experimental logic.

- **Result dimension:** Both approaches report improvement in surface quality, and the model further organizes results into predictive regimes, demonstrating its ability to generalize beyond individual parameter sets.

These observations indicate that the proposed framework can reconstruct human expert-level reasoning chains under fully automated conditions. By generating multiple sets of hypotheses consistent with real-world results, the system demonstrates both structural fidelity and creative generalization, rather than merely producing superficially plausible text. The complete evolutionary trajectory and analysis of the cryogenic SPDT case are presented in Appendix B. Further detailed examples are provided in Appendix C.

6 Conclusion

We formulate scientific hypothesis generation as a structured search problem over explicitly modeled reasoning spaces rather than unconstrained text generation. To this end, we propose the Hierarchical Scientific Reasoning Graph (HSRC-Graph) and Structure-Aware Hypothesis Evolution (SAHE), which jointly enable structure-preserving, evaluation-driven hypothesis evolution. Experiments on a large-scale materials science benchmark show that our approach produces hypotheses that are more logical, novel, feasible, and impactful than strong retrieval- and evaluation-based baselines. These results suggest that explicit reasoning structures and evolutionary search are key ingredients for controllable and interpretable automated scientific discovery.

Limitations

Despite the encouraging results, this work has several limitations that merit discussion.

First, the evaluation of generated hypotheses relies on an automated evaluation model with standardized criteria. While this enables scalable and consistent assessment across multiple dimensions, it cannot fully replace expert judgment, particularly for long-term scientific impact or domain-specific experimental feasibility. Future work could incorporate expert-in-the-loop evaluation or downstream experimental validation to further strengthen the assessment.

Second, due to the lack of mature and widely adopted baseline methods for automated scientific hypothesis generation in materials science, direct comparison with existing approaches remains inherently limited. Although we have implemented and evaluated MOOSE-CHEM, such methods were originally developed for the chemical domain. Additionally, rather than relying solely on potentially misleading direct comparisons, we place greater emphasis on controlled ablation studies, evolutionary trajectory analysis, and empirical experiments to isolate the contribution of each core component and to validate the effectiveness of structured representation and structure-aware evolutionary search.

Finally, the population-based evolutionary search introduces additional computational overhead compared to single-pass generation methods. Although this cost is justified by the observed gains in hypothesis quality and diversity, improving computational efficiency remains an important direction for future work.

References

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *Preprint*, arXiv:2506.13131.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Representation Learning*, volume 2025, pages 94003–94092.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2025. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723. Published online: 01 October 2025.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. *Preprint*, arXiv:2411.00816.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *Preprint*, arXiv:2411.02382.

Yinggan Xu, Hana Kimlee, Yijia Xiao, and Di Luo. 2025. Advancing ai-scientist understanding: Multi-agent llms with interpretable physics reasoning. *Preprint*, arXiv:2504.01911.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *Preprint*, arXiv:2410.07076.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of EMNLP Workshop of NLP for Science*.

A Appendix

A.1 Dataset Statistics and Distribution

Figure 5 presents an overview of the HSRC-5000 dataset, including the temporal distribution of publication years as well as the distribution of materials science subfields covered by the collected papers.The statistic in Figure B exceeding 100% is due to each paper being associated with multiple topics.

B Evolutionary Analysis of Idea Generation

B.1 Case Study: Cryogenic SPDT of PMMA

B.1.1 Original Paper Information

Paper Title: Machinability and Surface Properties of Cryogenic Poly(methyl methacrylate) Machined via Single-Point Diamond Turning

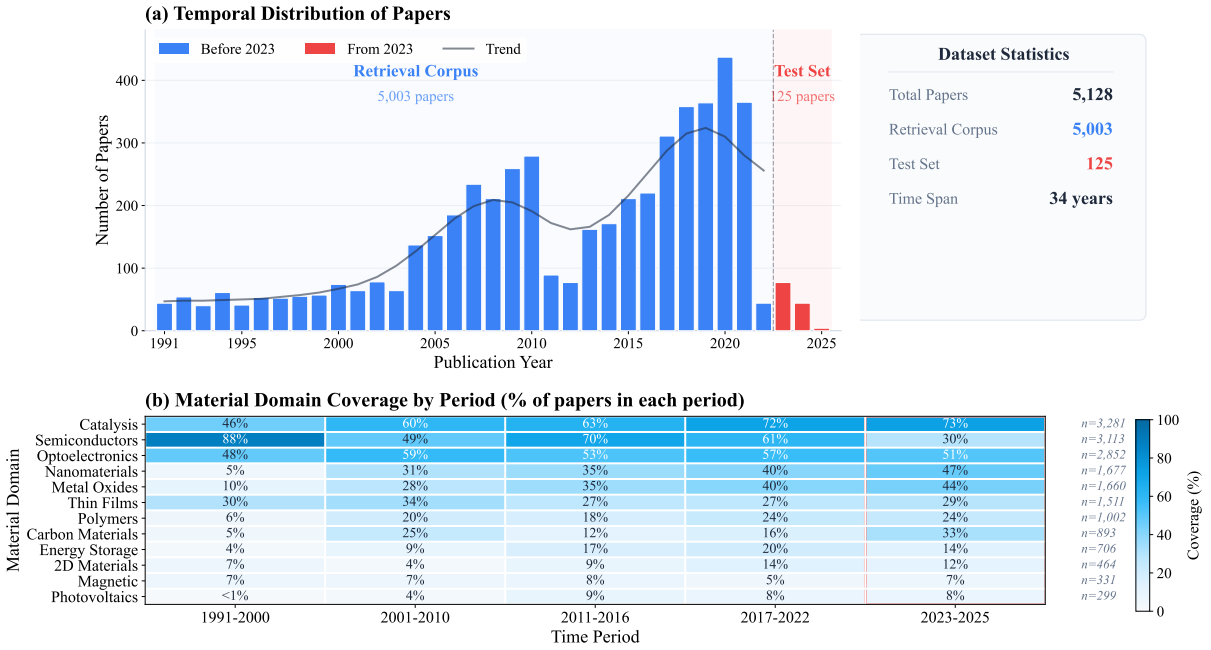

Figure 5: Overview of the HSRC-5000 dataset. (a) Temporal distribution of publication years and dataset splitting. (b) Distribution of materials science subfields covered by the papers.

Journal Information: *Materials* (Q2 Partition), Volume 17, Issue 4, Article 866, 2024

DOI: 10.3390/ma17040866

Paper Summary. Poly(methyl methacrylate) (PMMA), a representative optical polymer with a glass transition temperature exceeding 100 °C, is widely used in optical and precision engineering applications. Conventional single-point diamond turning (SPDT) of PMMA is typically conducted at ambient temperature, with research efforts primarily focusing on parameter optimization rather than deliberate modification of material properties.

The reference work explored the use of cryogenic cooling during SPDT of PMMA, demonstrating that lowering the machining temperature to 0 °C significantly enhances surface quality. Nanoindentation revealed increases of 37% in hardness and 22% in Young's modulus compared to room temperature. Taguchi optimization identified improved combinations of spindle speed, feed rate, and cutting depth under cryogenic conditions, yielding substantial reductions in surface roughness (Ra reduced from 11 nm to 6 nm) and profile height (Pt reduced from 291 nm to 124 nm). The study further established a mechanistic link between temperature, viscoelastic relaxation, mechanical properties, and surface formation through DMA data and a supporting theoretical model.

B.1.2 Evolutionary Trajectory Analysis

Idea Overview. **Idea ID:** eef8a904

Consistency Score: 8/10

Research Target. Develop and validate a temperature- and rate-normalized mechanical-intensity framework for cryogenic SPDT of PMMA. In its core form, the framework relies on a minimal descriptor set—mechanical intensity I_p, effective Deborah number $\mathrm{De}_{\mathrm{eff}}$, surface roughness/PSD, and simple birefringence or scattering metrics—to construct machinability and optical-safety maps. A higher-risk extension optionally incorporates spectroscopic descriptors (Raman, ellipsometry) aggregated into a single optical-disorder variable.

Positioning. This case study does not aim to reproduce or extend the original experiments directly. Instead, it serves as an illustrative example of how the proposed evolutionary idea-generation algorithm reconstructs, transforms, and refines a research concept when provided only with a high-level problem statement and heterogeneous seed knowledge.

B.1.3 Evolutionary Steps

Step 1: Initial Generation from a Cross-Domain Seed **Source Paper:** Physical and mechanical properties of Si:C:N films produced by remote microwave hydrogen plasma chemical va-

por deposition (*Applied Surface Science*, 2007)
DOI: 10.1016/j.apsusc.2007.03.027

Seed Contribution. The seed study systematically investigated how substrate temperature (30–400 °C) governs the composition, structure, and mechanical properties of Si:C:N thin films. Quantitative correlations were established between temperature, bonding configurations (Si–N, Si–C), and mechanical metrics such as hardness and modulus.

Mutation and Abstraction.

- **Temperature as a Primary Control Variable:** The explicit treatment of temperature as an independent tuning parameter inspired the elevation of workpiece temperature from a boundary condition to a first-class control variable in cryogenic SPDT.
- **Property–Process Mapping:** The compositional–mechanical mapping in thin films was abstracted into an analogous relationship between temperature, viscoelastic state, and machinability in PMMA.
- **Characterization Logic Transfer:** Multimodal characterization (XPS, FTIR) motivated the inclusion of polymer-specific structural probes (FTIR/Raman) to track chain conformation, physical aging, and surface chemistry changes induced by cryogenic conditioning.

Step 2: Method Recombination via a Dual-Control Analogy Source Paper: Dual dc-rf plasma deposited CN_x:H films with high elastic recovery (*Journal of Applied Physics*, 2007)
DOI: 10.1063/1.2402477

Recombination Strategy.

- **Dual-Control Mapping:** The separation of ion current density and ion energy in PECVD was reinterpreted as an orthogonal control of temperature (structural state) and mechanical intensity (loading rate and stress state) in machining.
- **Elastic Recovery as a Candidate Descriptor:** The prominence of elastic recovery in CN_x:H films motivated its elevation—alongside H/E and plasticity indices—as a candidate descriptor for balancing reversible and irreversible deformation in PMMA under high-rate contact.
- **Adapted Characterization Stack:** Techniques originally used for thin films (FTIR, Raman, XPS, profilometry, TEM) were selectively adapted to the polymer-machining context to interrogate near-surface structural evolution and subsurface damage.
- **Thin-Film Analogue Concept:** The notion of using PMMA films on rigid substrates as controlled analogues of the SPDT-affected surface layer emerged as a means to decouple thermal aging from cutting-induced deformation.

Step 3: Evaluation-Driven Refinement Systematic evaluation of the recombined idea exposed several weaknesses, triggering targeted refinements:

1. **Scope Compression and Phasing.** An initially over-extended instrumentation plan was restructured into phased stages, with a minimal must-succeed characterization set (DMA, nanoindentation, profilometry, force measurement) and advanced techniques reserved for confirmatory studies.
2. **Descriptor Validation.** Elastic recovery was reframed from an assumed governing parameter to a testable hypothesis. Rigorous indentation protocols and time–temperature superposition were introduced to align indentation timescales with SPDT contact mechanics.
3. **Controlled Use of Analogues.** Thin-film experiments were limited to isolating thermal densification and physical aging trends, with only qualitative validation against machined surfaces.
4. **Thermal Fidelity.** Nominal workpiece temperature was replaced by an effective cutting temperature metric, supported by thermal calibration, modeling, and environmental control.
5. **Statistical Rigor.** Formal design-of-experiments and replication were incorporated to ensure parameter efficiency and uncertainty quantification.
6. **Dimensional Stability.** Long-term form stability and residual stresses were explicitly included as evaluation criteria to prevent short-term surface gains from compromising optical performance.

B.1.4 Evolutionary Insights

Cross-Domain Knowledge Transfer. This trajectory illustrates how concepts from thin-film deposition—temperature-driven structural control, dual-parameter process spaces, and deformation descriptors—can be meaningfully transplanted into polymer machining.

Iterative Refinement. Evaluation acted as a selective pressure, pruning infeasible branches while preserving the core hypothesis structure, thereby transforming an exploratory concept into a disciplined, testable framework.

Innovation–Feasibility Balance. The final structure explicitly distinguishes between a conservative core path and higher-risk extensions, ensuring robustness without suppressing exploratory potential.

B.1.5 Final Research Framework

Core Path (Must-Succeed).

1. Define normalized mechanical-intensity descriptors (I_p, optionally I_W) and $\mathrm{De}_{\mathrm{eff}}$ using DMA and indentation-derived relaxation times.

2. Calibrate critical thresholds via scratch and orthogonal-cutting experiments with in-situ force measurement.

3. Construct machinability and optical-safety maps in I_p–$\mathrm{De}_{\mathrm{eff}}$–$T_{\mathrm{eff}}$ space using regularized statistical models.

4. Validate predictive capability on optical components using form, roughness, birefringence, and scatter measurements.

Optional Spectroscopic Branch. Spectroscopic descriptors are retained only if early-stage pilots demonstrate clear, monotonic correlations and measurable predictive gain over the mechanical-intensity baseline.

B.1.6 Expected Outcomes

The refined framework is expected to deliver:

1. Transferable procedures for estimating $\mathrm{De}_{\mathrm{eff}}$ and normalized intensity descriptors.

2. Descriptor-centered machinability and optical-safety maps with quantified uncertainty.

3. Cross-validated predictive models linking mechanical history to optical performance.

4. Generalizable design rules extensible to other optical polymers and precision machining processes.

B.1.7 Conclusions

This evolutionary case study demonstrates how cross-domain inspiration, when coupled with structured evaluation and refinement, yields research hypotheses that are both novel and implementable. The trajectory from thin-film deposition physics to cryogenic polymer machining exemplifies the intended behavior of the proposed idea-generation framework: reconstructing latent scientific logic, pruning infeasible paths, and converging toward robust, application-ready research programs.

C Detailed Case Analysis

C.1 Case Study 1: Plasmonic Sensor Optimization for Energy-Constrained Applications

C.1.1 Original Paper Information

Paper Title: Balancing detectivity and sensitivity of plasmonic sensors with surface lattice resonance

Journal Information: Nanophotonics (Q1 Partition), Volume 12, Issue 19, Pages 3721-3727, 2023

DOI: 10.1515/nanoph-2023-0225

Abstract: Resonators are at the core of optical sensors enhancing light–analyte interaction and leading to higher sensitivities. Maximizing the sensitivity is an obvious objective function for the resonator design. However, high sensitivity does not guarantee sufficient detectivity. When the optical energy budget is limited, as in sensors on mobile platforms, a higher sensitivity usually leads to lower detectivity for nanophotonic sensors. In such scenarios, resonator design requires balancing the trade-off between the sensitivity and detectivity of the resonant sensor. Here, we show the direct dependence of detectivity on the Q-factor and the trade-off between the Q-factor and sensitivity. We study this trade-off in an array of plasmonic resonators. We choose plasmonic resonators because of their high sensitivity arising from large local field enhancements. Then, we show that the detectivity of this sensor may be boosted for limited energy budget applications by making an array of resonators supporting a surface lattice resonance (SLR). We experimentally demonstrate sensing and detection of anti-mouse IgG protein in a gold nanodisk array–based SLR sensor for various energy budgets.

C.1.2 Generated Idea Analysis

Idea ID: a48d544f

Consistency Score: 9/10

Research Target Refine the project around explicitly defined, experimentally tractable, energy- and stability-normalized metrics, with a tightly scoped two-stage architecture.

Stage 1 (Primary, CMOS-compatible Objective): On a SiN/TiO_2 + Au SLR/Fano lattice platform interrogated by a low-power LED (less 50 μW at the chip), optimize three coupled metrics under mobile-relevant constraints:

1. **LOD_per_photon:** Minimum detectable anti-mouse IgG concentration at fixed LED power and fixed integration time, normalized by the number of photons arriving at the chip.

2. **LOD_per_joule:** Same LOD normalized by total energy consumed (LED power $\times$ time + detector bias power $\times$ time).

3. **Drift-/Response-Time-Normalized LOD:** Achievable repeatable LOD in a fixed window (e.g., 30 seconds), accounting for temperature drift and low-frequency noise.

Research Strategy Adopt an energy- and stability-centric co-design strategy that tightly couples optical, surface, packaging, detector, and electronic design:

1. **Metrics-First with System-Level Energy Budget:** Make LOD_per_photon, LOD_per_joule, and drift-/response-time-normalized LOD the organizing principle.

2. **Focused Stage-1 Process Space:** Constrain Stage-1 fabrication to a small, pre-justified set of variants.

3. **Integrated Noise and Packaging Strategy:** Explicitly treat surface/trap noise, packaging-induced drift.

4. **Stability-Aware Optical Design:** In Stage 1, co-opt intrinsic-loss engineering.

5. **Minimal but Informative Stage-2 Exploration:** Constrain Stage 2 to a small number of diamond resonator designs.

6. **Explicit Stage-1/Stage-2 Linkage:** Use the Stage-2 reference map to frame the Stage-1 results.

7. **Mobile-Like Validation and Electronics Co-Optimization:** Include at least one measurement campaign.

Research Method Augment and refine the existing seven-step method so that metrics, noise modeling, fabrication splits, and biosensing assays are explicitly tied to energy-normalized performance:

1. **Metrics, LOD Definition, and Hypotheses:** Define LOD as the lowest anti-mouse IgG concentration detectable with SNR $\geq$ 3; define LOD_per_photon = LOD / N_photons; define LOD_per_joule = LOD / E_total; quantify drift using Allan deviation and low-frequency noise PSD.

2. **Coupled-Mode, Noise, and System Modeling:** Extend coupled-mode models of SLR/Fano resonances; construct a hierarchical noise model (shot noise, thermal noise, surface/trap noise, environmental noise); validate model components experimentally.

3. **Stage-1 Fabrication Splits and Sample Strategy:** Fabricate a single, well-optimized SiN/TiO_2 + Au lattice geometry; implement three core process splits: A) baseline (no additional passivation beyond standard cleaning); B) one thin, conformal dielectric capping layer; C) one organosilane-based surface modification.

4. **Stability-Aware Optical Design:** In Stage 1, co-opt intrinsic-loss engineering to increase Q while maintaining reasonable sensitivity.

5. **Mobile-Oriented Interrogation and Electronics Energy Budgeting:** Use low-power LED ($\approx$ 50 μW); use low-bias silicon photodiode; explicitly measure total electrical energy consumption.

6. **Biosensing Chemistry and Microfluidics:** Use anti-mouse IgG as a model analyte; implement standard surface chemistry; include microfluidics to control sample delivery.

7. **Stage-2 Diamond + Detector Prototypes (Bounded Upper-Bound Study):** Fabricate a small number of diamond resonators; integrate detectors; measure LOD_per_photon and LOD_per_joule; create reference curves.

Expected Results With explicit metric definitions, a narrowed experimental design, and a unified noise and energy model, the revised project is positioned to deliver both quantitative performance gains and clear design rules for low-power, mobile-relevant biosensing.

Stage-1 Expectations (SiN/TiO_2 + Au, Aqueous):

- **Q-Factor:** An increase in aqueous Q for the SiN/TiO_2 + Au SLR/Fano lattice from $\approx$ 60–80 to $\approx$ 120–200 remains realistic.
- **Drift and Low-Frequency Noise:** A $\geq 2\times$ reduction in low-frequency noise PSD, with corresponding improvement in Allan deviation at relevant integration times (10–100 s).
- **LOD Improvements:** Under fixed LED power, a 2–3$\times$ reduction in both LOD_per_photon and LOD_per_joule.
- **System-Level Energy:** By explicitly measuring total electrical energy consumption (LED + detector), verify that the total energy budget is indeed within mobile-relevant limits (e.g., < 10 mJ per measurement).
- **Design Rules:** The combination of modeling and controlled process splits will yield explicit design rules, e.g., optimal dielectric capping thickness, surface modification chemistry, and packaging strategy.

Stage-2 Expectations (Diamond + Detectors, Upper Bound):

- Demonstration that ultra-high intrinsic Q (e.g., $>10^6$ in air) can translate to ultra-low LOD_per_photon under mobile-relevant energy budgets.
- Explicit reference curves relating (Q, detector bias, LED power) to LOD_per_photon and LOD_per_joule.
- Evidence that, when realistic aqueous biosensing and packaging constraints are considered, the Stage-1 SiN/TiO_2 + Au platform remains superior to the Stage-2 diamond platform in mobile-relevant scenarios.

Overall, the refined project should deliver not only measurable 2–3$\times$ reductions in LOD_per_photon and LOD_per_joule, but also clear design rules and benchmarks for low-power, mobile-relevant biosensing.

C.1.3 Consistency Analysis

This generated idea demonstrates high consistency with the original paper in the following aspects:

- **Research Problem:** Both focus on balancing sensitivity and detectivity in sensors under limited energy budgets, particularly for mobile platform applications. The original paper emphasizes optimizing the trade-off between these two factors, which aligns with the generated idea's focus on energy-normalized metrics, such as LOD_per_photon and LOD_per_joule.
- **Technical Approach:** Both employ surface lattice resonance (SLR) technology to enhance sensor performance under low-energy conditions. The original paper demonstrates the ability of SLR to balance sensitivity and detectivity, while the generated idea incorporates a similar strategy, with a more detailed focus on energy optimization, noise modeling, and multi-stage design.
- **Target Materials:** Both use gold nanodisk arrays as the base structure and explore optimization with dielectric materials like SiN/TiO_2. The material choice and integration with SLR technology are consistent across both the original paper and the generated idea, emphasizing the role of plasmonic resonators in achieving high sensitivity.
- **Solutions:** Both improve sensor performance by optimizing Q-factors and energy-normalized metrics (LOD_per_photon, LOD_per_joule). While the original paper focuses on optimizing the Q-factor to enhance detectivity, the generated idea goes a step further by incorporating energy normalization metrics and a two-stage experimental architecture, which further refines the performance improvement process.

Summary: The generated idea exhibits strong consistency with the original paper, retaining the core concept of balancing sensitivity and detectivity in plasmonic sensors under limited energy conditions. However, the generated idea introduces more detailed experimental metrics (such as LOD_per_photon and LOD_per_joule) and offers a multi-stage architecture for refining the sensor's energy efficiency. The new idea not only echoes

the original research's focus on SLR technology but also innovates by providing a clearer path for practical, energy-efficient sensor design in mobile applications.

C.2 Case Study 2: Nano-Cellular Topography for Enhanced Wear Resistance

C.2.1 Original Paper Information

Paper Title: Improving the Wear-Resistance of BT22 Titanium Alloy by Forming Nano-Cellular Topography via Laser-Thermochemical Processing

Journal Information: Materials (Q2 Partition), Volume 16, Issue 11, Page 3900, 2023

DOI: 10.3390/ma16113900

Abstract: This paper studies the microstructure, phase composition and tribological response of BT22 bimodal titanium alloy samples, which were selectively laser-processed before nitriding. Laser power was selected to obtain a maximum temperature just a little above the $\alpha \leftrightarrow \beta$ transus point. This allows for the formation of a nano-fine cell-type microstructure. The average grain size of the nitrided layer obtained in this study was 300–400 nm, and 30–100 nm for some smaller cells. The width of the "microchannels" between some of them was 2–5 nm. This microstructure was detected on both the intact surface and the wear track. XRD tests proved the prevailing formation of Ti_2N. The thickness of the nitride layer was 15–20 μm between the laser spots, and 50 μm below them, with a maximum surface hardness of 1190 HV0.01. Microstructure analyses revealed nitrogen diffusion along the grain boundaries. Tribological studies were performed using a PoD tribometer in dry sliding conditions, with a counterpart fabricated from untreated titanium alloy BT22. The comparative wear test indicates the superiority of the laser+nitrided alloy over the one that was only nitrided: the weight loss was 28% lower, with a 16% decrease in the coefficient of friction. The predominant wear mechanism of the nitrided sample was determined to be micro-abrasive wear accompanied by delamination, while that of the laser+nitrided sample was micro-abrasive wear. The cellular microstructure of the nitrided layer obtained after the combined laser-thermochemical processing helps to withstand substrate deformations and provide better wear-resistance.

C.2.2 Generated Idea Analysis

Idea ID: 6b776e23

Consistency Score: 8/10

Research Target Develop a practically scalable duplex surface-engineering framework for BT22 in which (i) near-transus processing plus compact plasma nitriding generate a nitrogen-rich nano-cellular diffusion zone with tunable boundary networks, and (ii) a low-complexity, modulus-graded $Ti_{1-x}Al_xN$-based coating overlays this substrate. Grain and column boundaries in both substrate and coating are treated as an explicitly engineered, nitrogen-enriched phase whose density, width, and chemistry can be adjusted within a bounded process window. The primary objective is to demonstrate, with minimal but well-contrasted conditions, that simple boundary-network descriptors (boundary area fraction, width, and connectivity proxies) can be linked to nitrogen transport, stress accommodation, and tribofilm stability, and used to design duplex architectures that sustain dry sliding and cyclic impact without delamination, excessive micro-abrasive wear, or friction-driven thermal damage, while preserving acceptable bulk mechanical and environmental performance of BT22.

Research Strategy Retain the duplex concept of a near-transus-processed and nitrided BT22 substrate plus a modulus-graded $Ti_{1-x}Al_xN$-based coating, but reshape the programme around staged, down-scoped exploration and simple, model-friendly boundary metrics. First, use a minimal factorial design with a few sharply contrasted nitriding and coating conditions to test whether engineered boundary regimes measurably influence tribology; only if clear trends emerge are additional conditions explored. Second, quantify boundary networks using a small, practical set of descriptors (boundary area fraction, mean boundary width, simple connectivity indicators from skeletonised images) and treat connectivity as a continuous parameter with uncertainty bands, rather than invoking strict percolation thresholds. These descriptors feed into a calibrated 1D lattice-plus-boundary diffusion model and simplified thermo-mechanical FE models that treat the boundary network as a softer, diffusion-enhancing phase, used to derive semi-quantitative design envelopes rather than exact predictions. Third, structure coating development in tiers: initially optimise Y-free TiN/$Ti_{1-x}Al_xN$ coatings and boundary states via process conditions alone, then introduce a single, carefully chosen Y level only after baseline behaviour is understood, using advanced characterisation on a small subset to verify where Y and nitrogen reside.

Throughout, introduce early, low-cost tribological screening gates before committing to deep STEM-EELS/ELNES/XANES or micro-mechanics, and embed simple checks on bulk mechanical integrity and high-temperature/oxidative response so that the boundary-centric design rules remain anchored to application-relevant performance.

Research Method

1. **Application Envelope and Decision Criteria:** Define representative dry sliding and cyclic-impact conditions (load, speed, temperature range, cycle count) and quantitative success metrics for friction stability, wear rate, and absence of cracking/delamination. Add boundary-centric targets expressed in simple metrics: minimum nitrogen-rich boundary area fraction or connectivity index in the diffusion zone and coating, and maximum allowable boundary width or hardness drop relative to grain interiors. Establish decision rules that prioritise the simplest viable architectures (nitrided BT22 plus a single-layer $Ti_{1-x}Al_xN$-based coating) whenever tribological and boundary metrics are met, reserving graded stacks or Y additions only to close clearly identified performance gaps.

2. **Near-Transus Processing and Baseline Integrity Checks:** Process BT22 near the alpha-beta transus to obtain an ultrafine alpha+beta microstructure and generate two surface states: as-ground and one down-selected water-confined LSP condition. Characterise near-surface grain size and boundary density via EBSD, and for one representative condition, confirm boundary width and morphology via TEM. Perform baseline mechanical and environmental checks (e.g. hardness, small-scale bend/fatigue indicators, and simple oxidation or high-temperature exposure tests) to ensure that near-transus processing and LSP do not unacceptably degrade bulk properties relative to application requirements.

3. **Minimal Nitriding DOE and Early Tribology Gate:** Conduct a compact plasma nitriding DOE on the two substrate states, limited initially to 2–3 well-separated conditions (e.g. low and high temperature plus one intermediate or one bias change) chosen to generate distinct nitrogen profiles and boundary architectures. For all conditions, measure nitrogen depth profiles (e.g. GDOES or SIMS), hardness/modulus gradients (nanoindentation), and residual stress (XRD). On only 2–3 contrasting conditions (e.g. shallow vs deep diffusion, low vs high boundary density), run early tribological screening tests (short-duration dry sliding and simple cyclic-impact) to identify promising regimes before committing to full characterization.

4. **Focused Coating Development and Boundary-State Mapping:** Deposit $TiN/Ti_{1-x}Al_xN$ coatings (initially Y-free) on selected nitrided substrates using a small, structured matrix of deposition parameters (Al content, bias, temperature) chosen to span a range of modulus and residual stress. Characterise coating microstructure (grain size, column width, texture) and boundary chemistry (via EDS, TEM on a subset). For a few key coating conditions, measure nitrogen diffusion through the coating into the substrate (SIMS depth profiling) and correlate with coating boundary descriptors. Introduce a single, carefully chosen Y level only after baseline TiN/TiAlN behaviour is understood, using advanced characterisation (STEM-EELS, APT) on a small subset to verify where Y and nitrogen reside.

5. **Integrated Modeling and Design Envelope Derivation:** Develop a calibrated 1D lattice-plus-boundary diffusion model for nitrogen in BT22 (and in representative coated stacks) that uses experimentally derived boundary metrics as inputs. Run simplified thermo-mechanical FE models of contact under dry sliding to explore how modulus gradients, residual stress, and thermal conductivity (influenced by boundary networks) affect interfacial stresses and temperature rise. Combine modeling results with experimental tribological data to derive semi-quantitative design envelopes for acceptable ranges of elastic-modulus mismatch, residual stress magnitude and gradient, coating thickness, and boundary-network intensity.

6. **Validation and Design-Rule Consolidation:** Perform full tribological validation (extended dry sliding, cyclic impact, and thermal cycling) on a small number of carefully selected

architectures that span the design envelopes. Use post-mortem analysis (cross-sectional SEM/TEM, EBSD, Raman) to identify failure modes (cracking, delamination, micro-abrasive wear) and correlate with boundary metrics and model predictions. Consolidate findings into explicit design rules linking boundary-network descriptors to nitrogen transport, stress accommodation, and tribofilm stability, with uncertainty bands and clear go/no-go criteria for different application scenarios.

Expected Results

1. A boundary-centric, yet practical, design framework for duplex BT22 architectures that specifies acceptable ranges of elastic-modulus mismatch, residual stress magnitude and gradient, coating thickness, and boundary-network intensity (expressed via boundary area fraction, mean boundary width, and simple connectivity indices), together with explicit uncertainty bands. These envelopes will be grounded in a combination of experiments and simplified models and validated against observed cracking, micro-abrasive wear, and delamination paths under dry sliding and cyclic impact.

2. A mechanistic comparison of a small number of carefully contrasted substrate/diffusion architectures, showing how grain size, grain-boundary density, nitrogen penetration depth, and nano-cellular/boundary-network descriptors influence crack initiation, subsurface damage evolution, adhesion, and the stability of nitrogen-rich networks once part of the coating is worn. This comparison will explicitly include checks that bulk mechanical indicators (e.g. simple fatigue surrogates) and oxidation behaviour remain within acceptable bounds.

3. A compact process–composition–microstructure map for TiN/$Ti_{1-x}Al_xN$ (and, where beneficial, a single low Y level) that links Al content, Y addition, deposition temperature, and bias to hardness/modulus, residual stress, and a small set of boundary descriptors (boundary density, width, and segregation tendency). The map will highlight a narrow, practically accessible window where coatings offer robust mechanical properties while exhibiting boundary states that support nitrogen retention/transport and tribofilm stabilisation, without excessive softening or loss of adhesion.

4. A calibrated 1D lattice-plus-boundary diffusion model for nitrogen in BT22 (and in representative coated stacks) that uses experimentally derived boundary metrics as inputs. While not attempting to fully capture 3D percolation, the model will clarify how much additional nitrogen penetration depth is realistically achievable through engineered boundary networks, and will provide a tool for extrapolating from the limited experimental conditions to broader process windows.

5. Demonstrated tribological performance improvements (e.g., $\geq$20% reduction in wear rate, $\geq$10% reduction in friction coefficient, and elimination of delamination) for at least one duplex architecture relative to single-stage nitriding, with clear links to boundary-network engineering and model-predicted design envelopes.

C.2.3 Consistency Analysis

The generated idea demonstrates strong consistency with the original paper in the following aspects:

- **Research Problem:** Both address the challenge of improving wear resistance of BT22 titanium alloy through surface modification techniques. The original paper uses laser-thermochemical processing to enhance wear resistance by forming nano-cellular topography, which is directly mirrored in the generated idea's approach of using near-transus processing and nitriding.

- **Technical Approach:** Both employ near-transus laser processing combined with nitriding to create nano-cellular microstructures. While the original paper focuses on the formation of a nano-cellular structure for wear resistance, the generated idea expands on this by adding a duplex coating strategy, with a focus on modulating the boundary network to further improve performance.

- **Target Materials:** Both focus on BT22 titanium alloy and nitrogen-based surface treatments. The original paper emphasizes the role

of Ti_2N formation and nitrogen diffusion in enhancing the microstructure, while the generated idea builds on this by introducing a $Ti_{1-x}Al_xN$-based coating, offering an additional layer of wear resistance.

- **Solutions:** Both aim to create nano-cellular topography with enhanced nitrogen diffusion to improve tribological performance. While the original paper demonstrates the effectiveness of laser+nitrided BT22, the generated idea innovates by engineering the boundary networks and developing a more sophisticated coating strategy that explicitly addresses tribological challenges.

Summary: The generated idea is highly consistent with the original paper in addressing the same research problem and applying similar technical approaches, such as laser-thermochemical processing and nitriding for wear resistance. However, the generated idea extends the original concept by introducing a duplex coating strategy and a more in-depth focus on boundary engineering, which aims to further enhance wear resistance and tribological performance. While the original research is limited to laser processing and nitriding, the generated idea adds a new layer of innovation by combining these treatments with advanced coating development and detailed modeling to optimize performance.

C.3 General Summary and Analysis of Differences and Similarities

In both case studies, the generated ideas demonstrate strong consistency with the original papers, reflecting a deep understanding of the core research concepts and approaches. Both generated ideas build on the foundation laid by the original studies, maintaining the same objectives and using similar materials or technologies. However, they introduce significant improvements and novel additions, including the introduction of more precise metrics, multi-stage experimental designs, and advanced modeling techniques.

Key differences between the generated ideas and the original research include:

- **Level of Detail:** The generated ideas incorporate more detailed experimental frameworks, including new energy normalization metrics (e.g., LOD_per_photon, LOD_per_joule) and multi-stage architectures, providing a clearer path for optimizing the performance of the sensors and wear-resistant materials.

- **Novelty in Solutions:** While the original papers focus on specific treatments or materials (e.g., SLR or laser processing), the generated ideas innovate by integrating additional elements such as boundary network engineering and duplex coatings, enhancing the overall functionality and performance.

- **Modeling and Simulation:** The generated ideas place a stronger emphasis on modeling and simulation to predict performance outcomes, such as using a 1D lattice-plus-boundary diffusion model in the wear resistance case or employing energy- and stability-centric co-design in the plasmonic sensor case.

Summary: In summary, the generated ideas are closely aligned with the original papers, but they extend the research by incorporating additional layers of complexity, optimization, and predictive modeling, resulting in a more comprehensive approach to addressing the challenges posed in both fields.

C.4 Case Study 3: Active Optical Lenses for Arc Flash Detection

C.4.1 Original Paper Information

Paper Title: Development of an Active Optical Lens for Arc Flashing Detection

Journal Information: Sensors (Q2 Partition), Volume 23, Issue 5, Page 2629, 2023

DOI: 10.3390/s23052629

Abstract: This paper contains the design of active optical lenses used for the detection of arc flashing emissions. The phenomenon of an arc flashing emission and its characteristics were contemplated. Methods of preventing these emissions in electric power systems were discussed as well. The article also includes a comparison of commercially available detectors. An analysis of the material properties of fluorescent optical fiber UV-VIS-detecting sensors constitutes a major part of the paper. The main purpose of the work was to make an active lens using photoluminescent materials, which can convert ultraviolet radiation into visible light. As part of the work, active lenses with materials such as Poly(methyl 2-methylpropenoate) (PMMA) and phosphate glass doped with lanthanides, such as terbium (Tb^{3+}) and europium (Eu^{3+}) ions, were analyzed. These lenses were used to make optical

sensors, which were supported by commercially available sensors in their construction.

C.4.2 Generated Idea Analysis

Idea ID: 2f920415

Consistency Score: 8/10

Research Target Develop strain- and codoping-engineered Ce^{3+} (later Pr^{3+})-activated $Gd_2Zr_2O_7$ and $Gd_2Hf_2O_7$ ceramics that act as UV-to-visible converting lenses or coatings for fast, robust arc-flash detection. Material specifications (bandgap, PL spectrum, decay time, afterglow, transparency) are derived from measured arc spectra and device-level modeling, and the project delivers transferable design rules for $A_2B_2O_7$ oxides linking codoping, oxygen-defect chemistry, strain, microstructure, and optical performance.

Research Strategy Use a tightly scoped, application-driven, staged program: (1) characterize real arc-flash UV spectra and temporal profiles and model simple detector modules to back-calculate material targets; (2) select a single primary host ($Gd_2Zr_2O_7$) and a small number of codopant pairs, guided by modest DFT calculations and design-of-experiments planning; (3) build a focused defect–strain–microstructure–optical map using a prioritized core characterization set, with advanced probes applied only to a few key compositions; (4) in parallel, develop both bulk semi-transparent ceramics and thin-plate or coating embodiments to de-risk the transparent-lens requirement; and (5) integrate the best materials into prototype arc-flash sensor modules and benchmark against commercial devices, iteratively refining the design rules and extending them to $Gd_2Hf_2O_7$ and Pr^{3+} only after clear trends are established.

Research Method

1. **Phase 0A – Application and Benchmarking Specification:** Measure or compile realistic arc-flash spectra (spectral power from roughly 200–400 nm, pulse duration, repetition characteristics, distance dependence) and typical ambient backgrounds in relevant switchgear environments. Model simple lens or coating plus photodiode or CMOS detector stacks to back-calculate material targets: required UV absorption band and cross section, PL emission band matched to detector responsivity and optical filters, maximum acceptable PL decay time and slow-tail fraction, acceptable afterglow and radiation-induced darkening, and minimum transmission at key UV and visible wavelengths for different thicknesses. Benchmark commercial arc-flash UV detectors and standard phosphor or lens materials (e.g. representative aluminates, silicates, fluorides, or commercially used UV converters) on the same testbed to quantify the performance gap and justify focusing on $Gd_2Zr_2O_7$-type hosts.

2. **Phase 0B – Host Justification and Scope Reduction:** Based on literature and initial screening, select $Gd_2Zr_2O_7$ as the primary host for detailed study, with $Gd_2Hf_2O_7$ reserved for a later validation or extension phase. Use simple comparative tests or literature meta-analysis to confirm that $Gd_2Zr_2O_7$ offers a credible combination of radiation tolerance, UV transparency, and compatibility with Ce^{3+} or Pr^{3+} activators relative to simpler hosts. Define Ce^{3+} as the primary activator for fast 4f–5d emission, with Pr^{3+} introduced only if needed for further lifetime optimization once the codoping and strain strategy is validated.

3. **Phase 1 – Modeling-Guided Dopant Selection and DOE Plan:** Use modest-scale DFT or hybrid-DFT calculations on $Gd_2Zr_2O_7$ to evaluate formation energies and preferred sites for Ce^{3+} on the Gd sublattice and for a small set of aliovalent dopants such as Ca^{2+} or Sr^{2+} (A-site) and Nb^{5+} or Ta^{5+} (B-site). Compute, at least qualitatively, associated oxygen-vacancy formation tendencies, local lattice strain, and bandgap modifications. Use these results to guide a focused experimental DOE that systematically varies codopant type and concentration, Ce^{3+} content, and processing conditions (sintering temperature, atmosphere) to map out defect–strain–microstructure space.

4. **Phase 2 – Core Characterization and Defect–Strain–Microstructure–Optical Mapping:** For the DOE samples, prioritize a core characterization set: XRD for phase identification and lattice parameters (strain proxy), Raman for local disorder and defect signatures, UV–Vis for bandgap and color-center absorption, steady-state and time-resolved PL for efficiency and decay

kinetics, and SEM for microstructure. Apply advanced probes (TEM/STEM-EELS, XPS, thermogravimetric redox cycling) only to a few key compositions that represent extremes or promising regimes. Build a focused map linking codopant choice, oxygen-defect proxies, lattice strain, and microstructure to bandgap position, defect absorption tails, PL efficiency, time response, thermal robustness, and scattering.

5. **Phase 3 – Transparency and Coating Development:** In parallel, develop processing routes for both bulk semi-transparent ceramics (e.g., high-temperature sintering with atmosphere control, possibly HIP) and thin-plate or coating embodiments (e.g., screen printing, sol-gel, or pulsed laser deposition) to derisk the transparent-lens requirement. Define quantitative transparency targets (e.g., transmission at key UV/visible wavelengths and acceptable scattering losses) and include early go/no-go criteria if transparency proves too difficult. Document trade-offs among transparency, defect stability, residual strain, microstructure, and manufacturability.

6. **Phase 4 – Device Integration and Benchmarking:** Integrate the best-performing materials into prototype arc-flash sensor modules, incorporating appropriate optical filters and photodetectors. Benchmark against commercial arc-flash UV detectors under realistic test conditions (arc-flash-like UV pulses, ambient backgrounds, temperature cycling, mechanical stress). Measure key performance metrics: minimum detectable UV dose, temporal resolution, false-trigger rate, and sensitivity drift after accelerated aging. Use these results to iteratively refine the design rules and material targets.

7. **Phase 5 – Extension to $Gd_2Hf_2O_7$ and Pr^{3+}:** Only after clear trends are established in $Gd_2Zr_2O_7$ with Ce^{3+}, extend the validated design rules to $Gd_2Hf_2O_7$ and introduce Pr^{3+} as an alternative activator if needed for further lifetime optimization. Perform a focused validation set of experiments to confirm that the design rules transfer and identify any host- or activator-specific adjustments needed.

Expected Results The project is expected to deliver:

1. An application-calibrated set of material specifications for arc-flash UV-to-visible converters, including targeted absorption and emission bands, PL decay and afterglow limits, transparency targets, and acceptable stability windows derived directly from measured arc spectra and device modeling;

2. A focused, experimentally and computationally informed map for Ce^{3+}-activated $Gd_2Zr_2O_7$ that connects codopant choice, oxygen-defect proxies, lattice strain, and microstructure to bandgap position, defect absorption tails, PL efficiency, time response, thermal robustness, and scattering, with practical proxies based on XRD, Raman, UV–Vis, SEM, and simple thermogravimetric and XPS measurements;

3. Identification of a narrow combined oxygen-defect, codoping, and processing window, in a defect-fluorite-leaning structural regime, where UV transparency is high enough for the chosen device format, strain-tuned bandgaps avoid mid-gap color centers, and Ce^{3+} (and later, selected Pr^{3+}) emission shows fast decay (with most emission within tens of nanoseconds and minimal slow tail) and limited thermal quenching over the operating temperature range;

4. Robust processing and atmosphere-control recipes for both bulk semi-transparent ceramics and thin-plate or coating embodiments that reproducibly preserve the desired codoping and strain state through densification, with clearly documented trade-offs among transparency, defect stability, residual strain, microstructure, and manufacturability;

5. Device-level evidence, obtained under realistic arc-flash-like conditions, that codoped and strain-engineered $Gd_2Zr_2O_7$-based modules can meet or exceed key performance metrics of commercial arc-flash UV sensors, including lower minimum detectable UV dose at a fixed false-trigger rate, improved temporal resolution, and reduced sensitivity drift after accelerated aging;

6. General, application-grounded design rules for $A_2B_2O_7$ oxides linking codoping, oxygen-defect chemistry, strain, microstructure, and optical performance that can be extended to

other UV-to-visible conversion applications beyond arc-flash detection.

C.4.3 Consistency Analysis

This generated idea demonstrates strong consistency with the original paper in the following aspects:

- **Research Problem:** Both address the challenge of developing active optical elements for arc flash detection through UV-to-visible conversion.
- **Technical Approach:** Both employ photoluminescent materials doped with lanthanides to convert UV radiation to visible light.
- **Target Materials:** Both focus on rare-earth-doped materials (Ce^{3+}, Pr^{3+}, Tb^{3+}, Eu^{3+}) for UV-to-visible conversion.
- **Solutions:** Both aim to create efficient, fast-responding active lenses that can be integrated into optical sensor systems.

Summary: The idea extends the original research by introducing more sophisticated material systems ($A_2B_2O_7$ pyrochlores), comprehensive defect and strain engineering, and detailed device-level validation, demonstrating innovative development of the original active optical lens concept.