

# Materealize: a multi-agent deliberation system for end-to-end material design and synthesis

Seongmin Kim<sup>1\*</sup>, Jaehwan Choi<sup>1\*</sup>, Kunik Jang<sup>1\*</sup>, Junkil Park<sup>1,2</sup>, Varinia Bernales<sup>3,4,7</sup>, Alán Aspuru-Guzik<sup>3,4,5,6,7,8,9,10</sup>, and Yousung Jung<sup>1,2\*</sup>

\*Equal contribution <sup>1</sup>Department of Chemical and Biological Engineering (BK21 Four) and Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea <sup>2</sup>Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea <sup>3</sup>Department of Chemistry, University of Toronto, 80 St. George St., Toronto, ON M5S 3H6, Canada <sup>4</sup>Department of Computer Science, University of Toronto, 40 St George St., Toronto, ON M5S 2E4, Canada <sup>5</sup>Department of Materials Science & Engineering, University of Toronto, 184 College St., Toronto, ON M5S 3E4, Canada <sup>6</sup>Department of Chemical Engineering & Applied Chemistry, University of Toronto, 200 College St., Toronto, ON M5S 3E5, Canada <sup>7</sup>Acceleration Consortium, 700 University Ave., Toronto, ON M7A 2S4, Canada <sup>8</sup>Vector Institute for Artificial Intelligence, W1140-108 College St., Schwartz Reisman Innovation Campus, Toronto, ON M5G 0C6, Canada <sup>9</sup>Canadian Institute for Advanced Research (CIFAR), 661 University Ave., Toronto, ON M5G 1M1, Canada <sup>10</sup>NVIDIA Corporation, Santa Clara, CA, USA

Correspondence to: Yousung Jung [yousung.jung@snu.ac.kr](mailto:yousung.jung@snu.ac.kr)

## 1. Introduction

Recent advances in machine learning have enabled large-scale property prediction and inverse design for inorganic crystals, alongside data-driven predictors for stability/synthesizability and synthesis-recipe recommendation. [1,2] Yet a practical “discovery-to-realization” gap remains: these capabilities are fragmented across task-specific tools and assembling them into an end-to-end workflow still requires substantial domain knowledge and engineering effort. [3-8] Moreover, LLM-based synthesis assistants can produce fluent but chemically implausible procedures when not grounded in domain evidence. [9-11]

We aim to close this gap by (i) unifying heterogeneous domain tools into a single natural-language-driven workflow for end-to-end design and synthesis planning, and (ii) improving syn-

thesis planning quality beyond passive tool calling through structured, tool-grounded scientific critique.

We introduce **Materealize**, a unified multi-agent system for end-to-end inorganic materials design and synthesis planning. *Materealize* combines (a) tool-level accuracy—via connected models for structure generation, property prediction, synthesizability prediction, and synthesis-recipe prediction—with (b) reasoning-level integration via a tool-grounded multi-agent debate protocol that refines synthesis procedures and mechanistic hypotheses.

## 2. System overview: Instant mode and Thinking mode

*Materealize* provides two complementary operation modes (Fig. 1), balancing latency and depth, offering distinct capabilities.

**Instant mode ( $\approx 1\text{--}2$  min per material).** Given a natural-language target (e.g., “bandgap  $\approx 1.34$

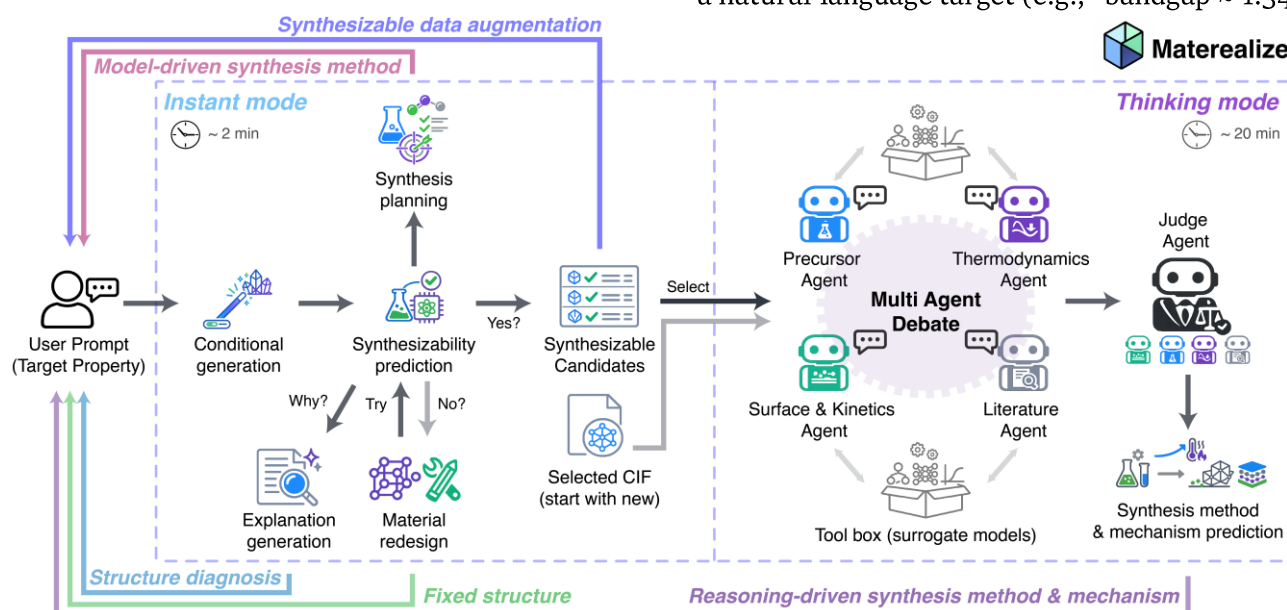


Fig. 1: Overall architecture of Materealize.

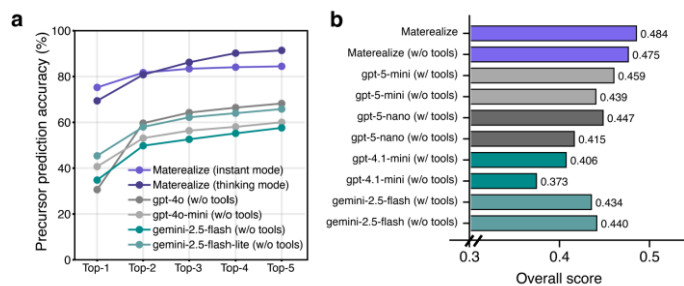
eV”), *Materealize* orchestrates connected tools to: (i) generate candidate crystal structures, (ii) evaluate synthesizability, (iii) filter unsynthesizable candidates, and (iv) predict synthesis recipes. Instant mode also supports diagnosis and redesign of unsynthesizable structures and unconditional synthesizable data augmentation. In our implementation, instant mode integrates 12 model endpoints spanning generative models, property prediction, synthesizability prediction, explanation, redesign, and synthesis-recipe prediction, along with crystallographic I/O and representation utilities.

**Thinking mode (~20 min).** *Materealize* addresses the accuracy limits of passive tool calling via a structured multi-agent debate (MAD) with four role-specialized agents (Precursor, Thermodynamics, Surface/Kinetics, and Literature Evidence) and a Judge that consolidates their arguments into a single synthesis report. The report provides recommended precursors/stoichiometry and a stepwise procedure and further predicts an accompanying mechanistic narrative such as key intermediates and the likely rate-limiting stage.

### 3. Evaluation and main results

#### Agent performance benchmark: recipes and mechanism-aware reports

We benchmark *Materealize* on two core synthesis-planning capabilities: (i) precursor (recipe) prediction accuracy and (ii) synthesis report quality that explicitly includes mechanistic reasoning (Fig. 2).



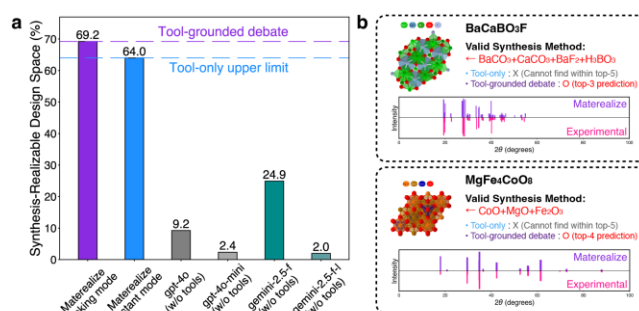
**Fig. 2: Performance benchmark for synthesis planning**

On a hold-out precursor benchmark, *Materealize* thinking mode consistently outperforms instant mode and tool-free agent baselines across all top- $k$  levels, achieving 86.2% (top-3), 90.2% (top-4), and 91.4% (top-5) accuracy (Fig. 2a). Beyond precursor selection, we evaluate the generated synthesis reports using an independent LLM-based evaluator over four axes—precursor, temperature, procedure, and mechanism—and find that *Materealize* attains the strongest overall performance (0.484 overall score) among the compared systems (Fig. 2b).

Collectively, these results indicate that tool-grounded deliberation not only improves recipe correctness but also yields higher-fidelity synthesis narratives, including mechanistic descriptions, compared to both tool-free reasoning and tool-connected single-model baselines.

#### Expanding the synthesis-realizable design space

To connect benchmark gains to end-to-end utility, we define a synthesis-realizable design space as the product of (i) the fraction of thermodynamically stable (or near-stable) generated structures and (ii) the probability of proposing a valid synthesis method within the top-5 suggestions. Under this metric, thinking mode expands the synthesis-realizable design space to 69.2%, surpassing the tool-only (instant) up-



per limit of 64.0% (Fig. 3a), while tool-free agent baselines remain substantially lower.

**Fig. 3: Tool-grounded debate expands the synthesis-realizable design space.**

#### Where debate helps: experimentally validated case studies

We further illustrate where the design space expands using experimentally validated examples (Fig. 3b). For compounds such as BaCaBO<sub>3</sub>F and MgFe<sub>4</sub>CoO<sub>8</sub>, tool-only execution fails to retrieve a valid synthesis route within the top-5 suggestions, whereas tool-grounded debate recovers valid routes within top-3/top-4, respectively. The associated diffraction comparisons (*Materealize* vs experimental) provide a qualitative anchor that the proposed routes are consistent with real synthesis outcomes (Fig. 3b). Together, these results support the central claim that *Materealize*'s tool-grounded, role-specialized deliberation improves not only recipe accuracy but also the mechanism-aware, experiment-facing quality of synthesis guidance, enabling a measurably broader region of materials that are both designable and synthesis-realizable.

## Acknowledgments

This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-MA2502-04. A.A.G. thanks Anders G. Frøseth for his generous support. A.A.G. and V.B. acknowledge the generous support of Natural Resources Canada and the Canada 150 Research Chairs program and the University of Toronto’s Acceleration Consortium, which receives funding from the CFREF-2022-00042 Canada First Research Excellence Fund.

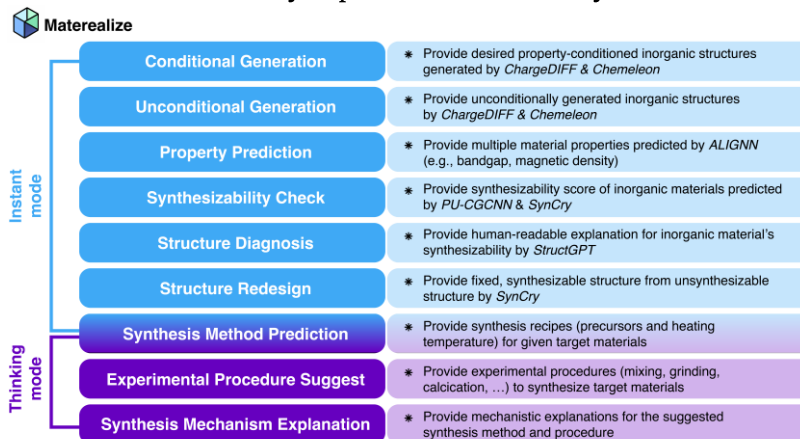
Reasoning. arXiv preprint arXiv:2509.23768 (2025).

## References

- [1] Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* 620, 47-60 (2023).
- [2] Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* 624, 80-85 (2023).
- [3] Zeni, C. et al. A generative model for inorganic materials design. *Nature* 639, 624-632 (2025).
- [4] Park, J., Choi, J. & Jung, Y. Integrating electronic structure into generative modeling of inorganic materials. arXiv preprint arXiv:2511.14228 (2025).
- [5] Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-based synthesizability prediction of crystals using partially supervised learning. *Journal of the American Chemical Society* 142, 18836-18843 (2020).
- [6] Choi, J., Kim, S. & Jung, Y. Synthesis-Aware Materials Redesign via Large Language Models. *Journal of the American Chemical Society* 147, 39113-39122 (2025).
- [7] Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials* 7, 185 (2021).
- [8] Kim, S., Schrier, J. & Jung, Y. Explainable Synthesizability Prediction of Inorganic Crystal Polymorphs Using Large Language Models. *Angewandte Chemie International Edition* 64, e202423950 (2025).
- [9] Kim, S., Noh, J., Gu, G. H., Chen, S. & Jung, Y. Predicting synthesis recipes of inorganic crystal materials using elementwise template formulation. *Chemical Science* 15, 1039-1045 (2024).
- [10] M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1-11 (2024).
- [11] Yang, C., Lu, J., Wan, H., Yu, J. & Qin, F. From What to Why: A Multi-Agent System for Evidence-based Chemical Reaction Condition

## Appendix A. Overview of *Materealize*'s core functional modules and capabilities.

In instant mode, *Materealize* can handle various tasks related to inorganic material design and synthesis within a few minutes. In thinking mode, *Materealize* can suggest more accurate and richer synthesis information within 20 minutes. We note that *Materealize* is designed to be highly extensible, making it easy to integrate new tools and seamlessly expand its functionality.



## Appendix B. Example of task capabilities by *Materealize*.

### (a) Task 1 : Conditional end-to-end design of synthesizable candidates

I want to find proper inorganic materials, which have bandgap around 1.34 eV

(Used tool sequences)

So, here is the **\*Synthesizable\*** candidates with your target property 1.34 eV:

And here is the **\*Synthesis Method\*** for each candidates as follows:

Top-1 method	Precursors: $K_2CO_3 + CaCO_3$ Heating temperature : 885.1°C
Top-2 method	Precursors: $K_2CO_3 + CaO$ Heating temperature : 921.5°C
⋮	

Completed time : ⌚ ≈ 4 min

### (b) Task 2 : Diagnosing and fixing unsynthesizable materials

I want to diagnose and fix this structure

(Used tool sequences)

It is **\*Unsynthesizable\*** structure by following reasons:

[Reason 1] : [Charge Imbalance]  
NaMnO2 cannot be synthesized because it has a charge imbalance.....

[Reason 2] : [Structural Instability]  
The structural configuration described suggests that the octahedral....

[Reason 3] : [Incompatible Oxidation States]  
The oxidation states of Na (+1) and Mn (+4) are incompatible...

And here is the fixed **\*Synthesizable\*** alternative:

Completed time : ⌚ ≈ 1 min

### (c) Task 3 : Detailed synthesis procedure and mechanism prediction for synthesizable candidates or user's materials

Please generate detailed synthesis procedure and possible reaction mechanism for **candidate "3"** or .... for this structure

(Running Multi-Agent Debate for candidate 3....) round 1.. 2 .. 3 .. 4 .. 5 consensus reached!

So, here is the detailed synthesis report for candidate 3,  $Ti_3Fe_5O_{12}$ :

**Synthesis Report (condensed version for readability)**

*Precursors & Stoichiometry*  
Solid-state synthesis from binary oxides:  $TiO_2 + Fe_2O_3$  (optional  $Fe_3O_4$  as a mild redox buffer if mixed-valence control is needed)

*Mixing / Mechanical Activation*  
Dry powders thoroughly, then ball-mill (wet with IPA or dry) to maximize mixing and reduce diffusion length scales. Dry and sieve to obtain a homogeneous fine powder before heat treatment ...

*Stepwise Heat Treatment (Solid-State Route)*  
(i) Pre-dry powders (~120 °C) to remove moisture/adsorbates.  
(ii) Pre-calcine in the 600–800 °C window to initiate interfacial reactions and form Fe–Ti oxide intermediates.  
(iii) Regrind + pelletize, then high-T anneal in the 900–1200 °C window (starting point ~1050 °C) to drive .....

*Atmosphere / Redox Control (Key Lever)*  
Oxygen chemical potential strongly affects phase evolution and impurity formation. Use flowing air/ $O_2$  as the default; apply  $O_2$  post-anneal (~600 °C) if excess  $Fe^{2+}$  is detected. If repeated secondary phases indicate redox mismatch, introduce a small  $Fe_3O_4$  buffer....

*Reaction mechanism*  
Reaction proceeds via interfacial mixing → binary titanate/spinel-type intermediates → long-range cation diffusion and reconstructive crystallization into  $Ti_3Fe_5O_{12}$  during high-temperature annealing. The rate-limiting step is bulk cation diffusion in the final conversion stage.

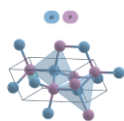
Completed time : ⌚ ≈ 20 min

## Appendix C. Example of multi-agent debate dialogue trace by *Materealize* thinking mode.

Given a synthesis target, *Materealize* starts from the precursor set proposed by the initial tool (ElemwiseRetro) and then conducts a structured debate in which specialized agents (precursor, thermodynamics, literature, and surface/kinetics) iteratively critique and refine the route, ultimately converging on an improved synthesis recommendation.

### A multi-agent debate dialogue flow

Synthesis target : AlP (aluminum phosphide)



Precursor Agent

Literature Agent

Judge Agent

Thermodynamics Agent

Surface & Kinetics Agent



ElemwiseRetro suggests a retrosynthesis pathway: a solid-state route using **AlN** with a P-containing precursor. AlN also appears in the surrogate outputs as an alternative precursor.

**I disagree with treating AlN as an equal primary option to elemental Al.** Converting AlN is kinetically and thermodynamically less straightforward, and it can introduce N contamination that is difficult to remove. Let's keep AlN as contingency only.



From what I can find in the literature, we should prioritize **elemental Al + red P at 1:1 molar stoichiometry**. **AlN + P is a plausible alternative**, but it would need strong validation—and I'm not seeing sufficient support to justify it as a primary route.

Practically, use **high-purity Al powder + red phosphorus** in a 1:1 stoichiometric ratio, with a slight P excess to offset P volatility. **Red P is safer and has lower volatility than white P.**



One more caution: in locally P-deficient regions, **Al-rich intermediates—notably Al<sub>3</sub>P<sub>5</sub>—can form**. If P activity is insufficient, kinetic trapping of Al<sub>3</sub>P<sub>5</sub> is likely. We should monitor for Al<sub>3</sub>P<sub>5</sub> during heating.

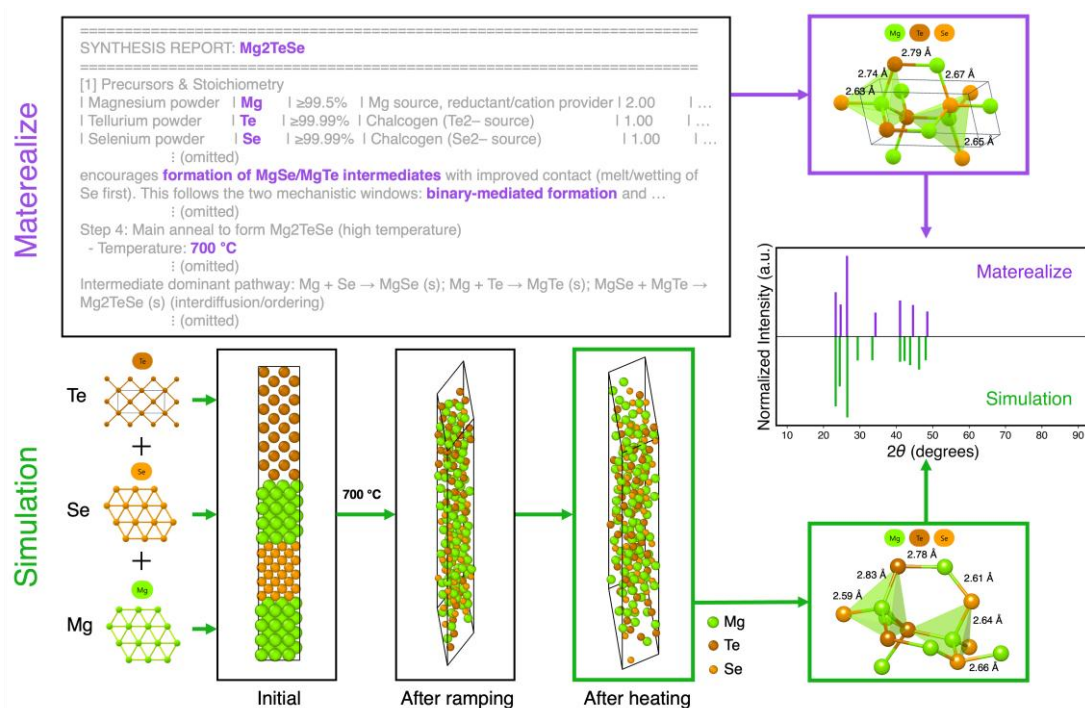
Agreed. We'll adopt **Al + red P (1:1 + slight P excess)** as the default precursor set. We'll reject AlN as a primary and keep it only as a contingency, and we'll explicitly note **Al<sub>3</sub>P<sub>5</sub> monitoring to avoid persistence of that intermediate**.



ML-suggested AlN is not used as a primary precursor. Final recommendation: **Al + red P (1:1 with slight P excess), with attention to P volatility and Al<sub>3</sub>P<sub>5</sub> as a potential transient intermediate**.

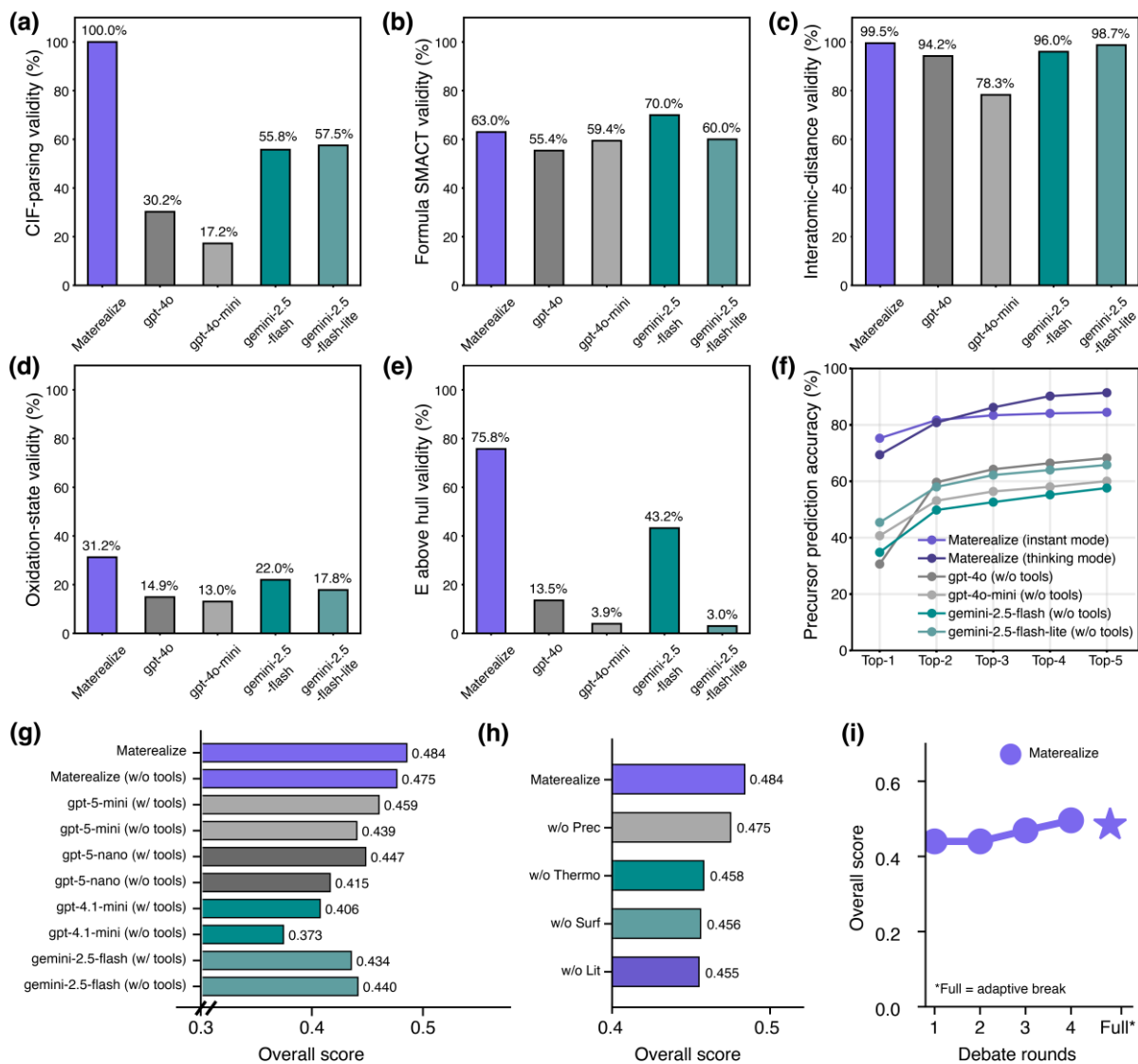
## Appendix D. Simulation-based validation of *Materealize* synthesis report.

For the Mg<sub>2</sub>TeSe target, *Materealize* generates a synthesis report and the corresponding target structure. We convert each report into a MACE-based molecular dynamics (MD) validation workflow by building an initial stacked precursor assembly (Initial), applying temperature ramping to the report-specified (700 °C) peak temperature (After ramping), and conducting isothermal annealing (After heating). We then compare powder XRD computed from the relaxed structure obtained by further optimizing the final MD snapshot (green) with the reference peaks from the target structure (purple). The simulated patterns reproduce the dominant target reflections.



**Appendix E. Extended figure for Fig 2. Agent performance benchmark.**

(a-e) Generated structure validity for (a) CIF-parsing, (b) formula SMACT check, (c) minimum interatomic-distance  $> 0.5 \text{ \AA}$ , (d) oxidation-state, and (e) energy above hull  $\leq 0.2 \text{ eV/atom}$  relative to MP references. (See also Extended Data Fig. 1 and 2 for element diversity and energy above hull distribution.) (f) Top-k exact match accuracy of suggested precursors. (g-i) Generated synthesis method and mechanism evaluation.



**Appendix F. Extended figure for Fig 3. Extending the synthesis-realizable space with *Materealize*.**

(a) Quantitative comparison of the synthesis-realizable design space across different agent systems. The synthesis-realizable design space is defined as the product of the thermodynamically stable ratio ( $E_{\text{hull}} \leq 0.2$  eV/atom) and the valid synthesis-method ratio within the top-5 predictions. (b) t-SNE visualization of materials space for which correct design-and-synthesis routes are identified by *Materealize* in both instant and thinking modes. Among the materials designed by *Materealize*, five experimentally-validated materials are shown, where tool-only execution fails to identify valid synthesis routes within the top-5 predictions, while thinking mode successfully proposes experimentally consistent synthesis methods.

