ONEMAT: GENERATIVE OPTIMIZATION FROM SCIENTIFIC LITERATURE FOR MATERIALS DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Materials design still relies heavily on costly physical experimentation and search over predefined candidate spaces, limiting both the efficiency and breadth of exploration. Recent generative approaches in materials science have advanced candidate-level design by proposing novel compositions and atomic structures, but executable synthesis procedures for their realization remain largely outside iterative generation and optimization. We introduce **OneMat**, a literature-driven framework for *synthesis-oriented generative optimization*, where LLMs generate and iteratively refine executable synthesis plans. To enable this paradigm, OneMat transforms scientific literature into a structured domain resource containing evidence-grounded synthesis knowledge, generation examples, and preference signals, and integrates generation with retrieval and evaluation in an iterative optimization loop. We evaluate OneMat on acidic oxygen evolution reaction (OER) catalyst design using a literature corpus of 5,642 papers across mining, synthesis-plan generation, optimization, and wet-lab validation. Evidence-first mining consistently reduces unsupported extraction across four LLMs, while literature-derived post-training improves synthesis-plan quality, increasing the evaluator score from 0.51 (Base) to 0.62 (SFT) and 0.69 (DPO), and enhancing closed-loop optimization. Finally, automated wet-lab experiments translate OneMat-generated synthesis plans into functional catalysts, with the best-performing design discovered through limited experimental exploration achieving an overpotential of 155 mV at 10 mA cm^{-2} and only a 21 mV increase after 500 CV cycles. Together, these results establish a route from scientific literature to experimentally validated synthesis-oriented generative optimization for materials design.

1 INTRODUCTION

Materials discovery requires navigating complex design spaces in which composition, structure, synthesis conditions, and processing procedures jointly determine material properties. For functional materials such as acidic oxygen evolution reaction (OER) catalysts, achieving high activity and stability requires balancing multiple objectives under experimentally demanding constraints (Kibsgaard & Chorkendorff, 2019; Seitz et al., 2016; Geiger et al., 2018; Seh et al., 2017; Nong et al., 2020; Roger et al., 2017). Conventional discovery workflows rely heavily on physical experimentation and predefined candidate spaces, limiting both exploration efficiency and access to unconventional design strategies. Recent generative approaches have expanded materials exploration by proposing novel compositions and atomic structures beyond predefined candidates, but they still primarily treat the material itself as the generation target rather than the executable synthesis procedures required for physical realization.

This motivates *synthesis-oriented generative optimization*, where executable synthesis plans themselves become optimization objects. Large language models (LLMs) provide a natural basis for this paradigm by acting as search operators that propose synthesis plans, incorporate feedback, and iteratively refine solutions beyond manually defined candidate spaces.

Despite recent progress in LLM-based scientific assistance, using LLMs as reliable optimization operators for synthesis-plan generation remains challenging. First, LLMs must generate scientifically grounded synthesis plans from heterogeneous literature, yet their outputs can contain plausible but unsupported statements originating from parametric memory rather than explicit source

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

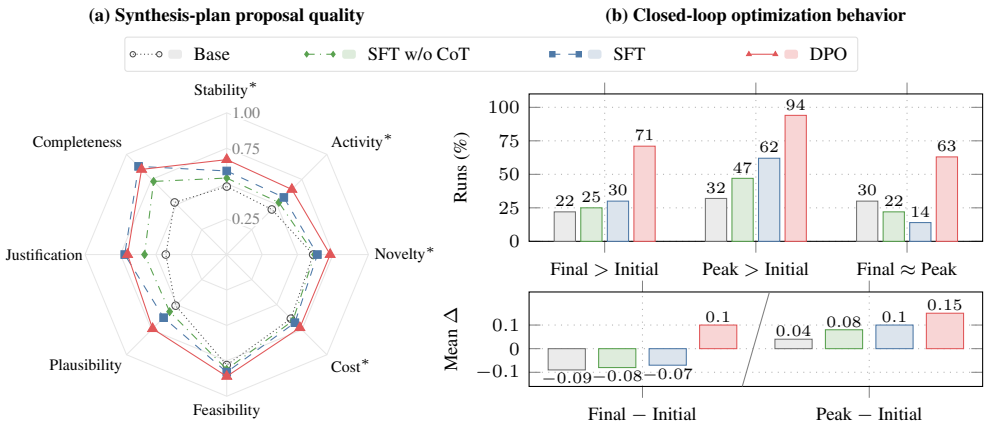


Figure 1: Synthesis-plan proposal quality and closed-loop optimization behavior. (a) Hybrid Evaluator profiles (0–1) of Base, SFT w/o CoT, SFT, and DPO on 180 held-out cases. Starred dimensions indicate the criteria emphasized in the overall evaluator score calculation, yielding scores of 0.51, 0.57, 0.62, and 0.69, respectively. (b) Closed-loop optimization behavior on the 100-case held-out batch. The upper plot reports the fraction of runs in which the final score exceeds the initial score, the within-run peak score exceeds the initial score, or the final score remains within 0.05 of the peak score. The lower plot reports mean changes in final and peak scores relative to initial scores.

evidence (Maynez et al., 2020; Ji et al., 2023). Existing approaches such as retrieval augmentation (Lewis et al., 2020), chain-of-thought prompting (Wei et al., 2022), and post-generation verification (Dhuliawala et al., 2024) mainly improve reliability during inference, whereas constructing reliable scientific supervision directly from literature remains less explored. Second, scalable and reliable preference signals are needed to guide optimization toward experimentally supported solutions. Current alignment methods rely on human feedback (Christiano et al., 2017; Ouyang et al., 2022) or model-generated judgments (Bai et al., 2022; Lee et al., 2024), but expert annotation is costly and evaluator-based approaches depend on external models. Third, systematic iterative optimization of executable synthesis plans remains challenging. Although LLMs have shown potential for scientific hypothesis generation (Zhou et al., 2024; Abdel-Rehim et al., 2025) and feedback-driven refinement (Madaan et al., 2023), existing approaches are not designed to optimize structured experimental procedures under domain-specific evaluation signals. Together, these challenges limit the use of LLMs as reliable synthesis-plan optimization agents.

To address these challenges, we introduce **OneMat**, a literature-driven framework for *synthesis-oriented generative optimization* that treats scientific literature as both a knowledge source and a source of optimization supervision. Scientific papers provide two complementary signals: explicit evidence describing how materials are synthesized and comparative experimental results indicating which design choices perform better under matched conditions. OneMat transforms these signals into evidence-grounded synthesis knowledge, generation and reasoning supervision, and experimentally derived preference pairs for adapting LLM-based generators before iterative synthesis-plan optimization. Specifically, **Evidence-Grounded Mining (EGM)** constructs synthesis supervision through evidence-first structured extraction to reduce unsupported extraction, while **Literature Preference Learning (LPL)** derives scalable preference signals from experimentally reported comparisons without costly expert annotation or external judge models. Finally, a feedback-driven optimization loop combines generation, domain-specific evaluation, and evidence retrieval to iteratively refine executable synthesis plans before experimental validation.

We instantiate OneMat on acidic OER catalyst design using 5,642 literature records and evaluate it across literature mining, synthesis-plan generation, closed-loop optimization. The resulting synthesis plans are further realized through automated wet-lab experiments, providing prospective validation of the framework.

Our contributions are summarized as follows:

- **A framework for synthesis-oriented generative optimization.** We introduce **OneMat**, which extends generative materials design beyond candidate-level generation by treating executable synthesis plans as optimization objects and using scientific literature to support their iterative generation and refinement.
- **Evidence-grounded dataset construction and literature-derived supervision.** We introduce **Evidence-Grounded Mining (EGM)** to construct a structured acidic OER catalyst dataset with 5,642 literature records and explicit source evidence. The resulting resource yields 3,411 retrieval entries, 29,148 supervised examples, and 1,012 experimentally grounded preference pairs. **Literature Preference Learning (LPL)** uses these reported comparisons for preference alignment without costly expert annotation or external judge models.
- **Computational and experimental validation.** We show that literature-derived supervision improves synthesis-plan generation and closed-loop refinement, and prospectively validate ONEMAT-generated plans through an automated wet-lab workflow, culminating in a physically realized RuO_2/MnO_2 catalyst.

2 BACKGROUND AND RELATED WORK

2.1 GENERATIVE MATERIALS DESIGN AND SYNTHESIS-AWARE GENERATION

Recent generative models have expanded materials exploration by directly generating novel compositions and atomic or crystal structures, including diffusion-based crystal generation (Jiao et al., 2023; Zeni et al., 2025) and flow-based modeling of periodic materials (Miller et al., 2024). Emerging work has also begun to incorporate synthesizability, for example by jointly modeling molecular structures and reaction pathways (Rekesh et al., 2026). These methods primarily operate at the composition, molecular, or atomic-structure level, whereas ONEMAT treats executable synthesis procedures as objects of iterative generation and optimization.

2.2 LITERATURE-GROUNDED SCIENTIFIC SUPERVISION AND PREFERENCE LEARNING

Scientific information extraction has progressed from text-mined synthesis databases and materials-specific extraction systems (Kononova et al., 2019; Mysore et al., 2019; Swain & Cole, 2016; Kim et al., 2017) to LLM-based methods for processing heterogeneous scientific documents and constructing materials datasets (Schilling-Wilhelmi et al., 2025; Ansari & Moosavi, 2024; Chen et al., 2025). Retrieval-augmented and self-reflective generation can improve source grounding during inference (Lewis et al., 2020; Asai et al., 2024), while FActScore (Min et al., 2023) emphasizes claim-level support. In parallel, preference-based adaptation commonly relies on human feedback (Christiano et al., 2017; Ouyang et al., 2022), DPO (Rafailov et al., 2023), or AI-generated and self-rewarding signals (Bai et al., 2022; Lee et al., 2024; Yuan et al., 2024), but remains sensitive to judge quality and bias (Zheng et al., 2023). In contrast, ONEMAT derives both grounded supervision and preference signals directly from experimentally reported literature evidence.

2.3 LLM-BASED SCIENTIFIC AGENTS AND CLOSED-LOOP OPTIMIZATION

Materials discovery and optimization increasingly combine computational screening, active learning, Bayesian optimization, and autonomous experimentation (Merchant et al., 2023; Lei et al., 2025; Lookman et al., 2019; Shahriari et al., 2016; Szymanski et al., 2023; Burger et al., 2020; Tabor et al., 2018). In parallel, LLM-based scientific systems have demonstrated capabilities in property prediction, chemistry tool use, synthesis planning, reasoning, and autonomous or multi-agent research workflows (Jablonka et al., 2024; Bran et al., 2024; Boiko et al., 2023; Bran et al., 2026; Zheng et al., 2025; Wang et al., 2025; Hu et al., 2025; Ghafarollahi & Buehler, 2025). Unlike classical optimization over predefined candidate or experimental spaces, ONEMAT uses an LLM-based generator to construct and iteratively refine executable synthesis plans under domain-specific evaluation and retrieved literature evidence.

3 SYNTHESIS-ORIENTED GENERATIVE OPTIMIZATION

ONEMAT realizes synthesis-oriented generative optimization through a three-stage pipeline from scientific literature to executable synthesis-plan refinement. First, EGM transforms heterogeneous literature into an evidence-grounded domain resource containing structured synthesis knowledge, generation and reasoning supervision, and experimentally derived preference pairs. Second, LPL adapts an LLM-based generator through supervised fine-tuning and literature-derived preference alignment. Third, the agentic optimizer integrates the adapted generator with retrieval and domain-specific evaluation in an iterative evaluate–retrieve–revise loop, where the generator acts as a search operator over executable synthesis plans rather than selecting from a predefined candidate space. Figure 2 illustrates the overall pipeline.

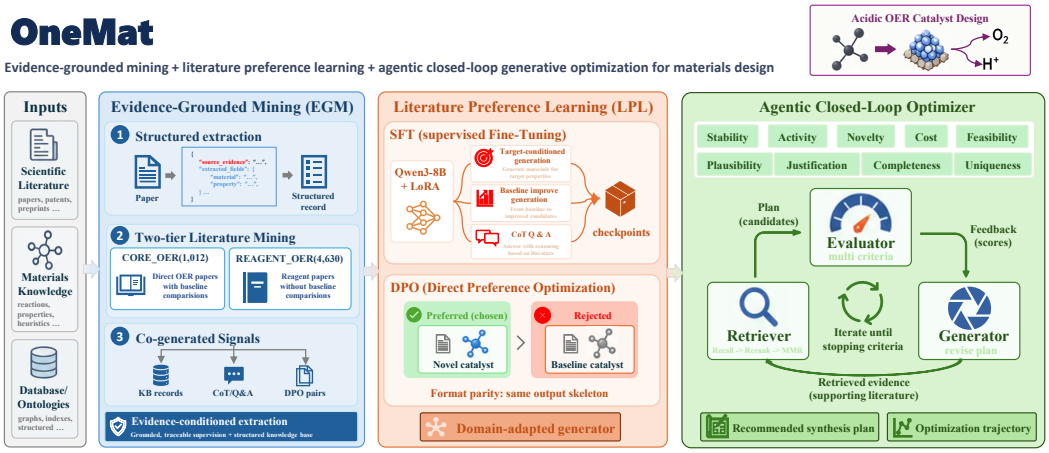


Figure 2: Overview of ONEMAT. EGM constructs literature-grounded data and supervision, LPL adapts the generator, and an evaluate–retrieve–revise loop performs synthesis-plan optimization before experimental validation.

3.1 STAGE 1: EVIDENCE-GROUNDED MINING

Evidence-first structured extraction. Given a paper p, EGM performs structured extraction $Y = F(\text{prompt}, p)$ into a predefined JSON schema (Appendix A). We place the verbatim `source_evidence` field before extracted values, creating an evidence-first autoregressive dependency in which attributes are generated only after explicit textual evidence has been identified. By contrast, generating evidence after extracted content may allow plausible values to arise from parametric knowledge before source grounding, increasing the risk of unsupported extraction (Maynez et al., 2020). This ordering therefore strengthens the connection between extracted claims and their supporting evidence.

OER catalyst dataset construction and characterization. Applying EGM to the acidic OER literature produced 5,642 structured records in two complementary tiers. **CORE_OER** contains 1,012 comparative records with matched novel–baseline catalyst pairs for preference learning, while **REAGENT_OER** contains 4,630 synthesis-focused records without matched comparisons, broadening coverage of synthesis procedures, reagents, and processing conditions for domain adaptation and retrieval. The resulting dataset captures broad synthesis and catalyst-composition spaces, spanning diverse preparation strategies and material compositions. Figure 3 summarizes the extracted dataset landscape, including comparative performance signals, synthesis, reagent, elemental, and processing-condition spaces.

Retrieval knowledge and supervision construction. Based on the structured dataset generated by EGM, we construct three complementary resources for retrieval, generator adaptation, reasoning, and preference learning:

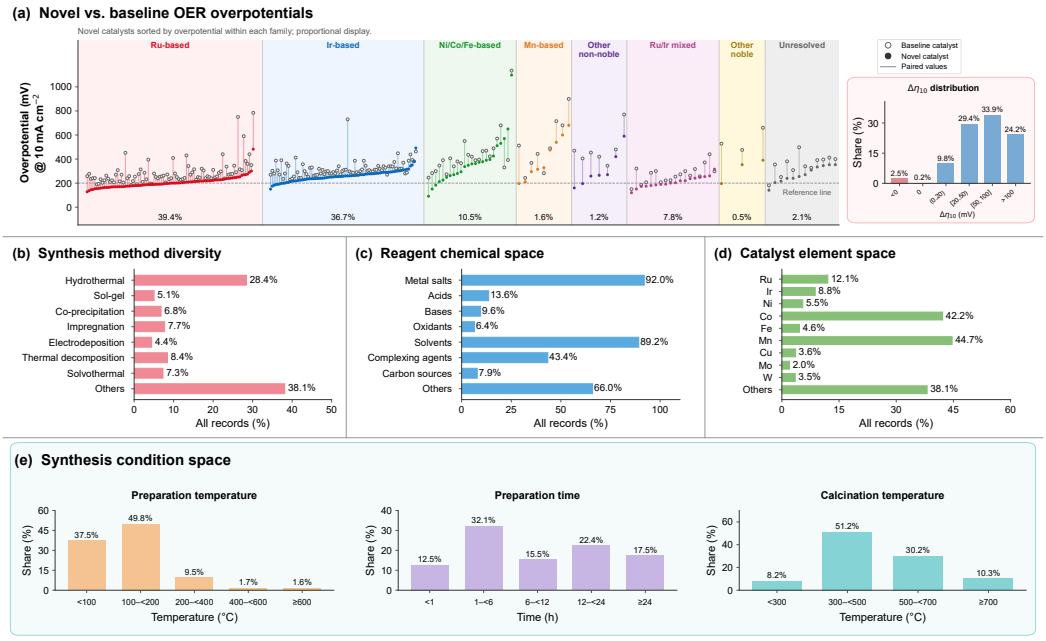


Figure 3: Characterization of the acidic OER catalyst dataset. (a) Comparative performance of novel and baseline catalysts in CORE_OER, grouped by the element family of the novel catalyst. The improvement signal is defined as $\Delta\eta_{10} = \eta_{10}^{\text{baseline}} - \eta_{10}^{\text{novel}}$, with positive values indicating improvement. The dashed line marks 200 mV, and the inset shows the distribution of $\Delta\eta_{10}$. Points show a stratified subset to reduce overlap. (b)–(d) Coverage of synthesis methods, reagent chemical space, and catalyst-element composition, respectively. (e) Distributions of preparation temperature, preparation time, and calcination temperature for records with valid single-value entries. Percentages in multi-label panels represent record coverage and are not expected to sum to 100%.

- **Retrieval knowledge records**: structured catalyst, synthesis, performance, and evidence fields used to build the vector knowledge base for evidence-grounded retrieval.
- **Generation and reasoning examples**: target-conditioned synthesis-plan examples, baseline-to-novel transformation examples from comparative records, and evidence-grounded question–answer pairs derived from source context.
- **Preference pairs**: comparative novel–baseline examples, with the reported novel catalyst as the chosen response and the baseline catalyst as the rejected response.

The resulting resources contain 3,411 unique retrieval knowledge entries, 29,148 supervised fine-tuning examples, and 1,012 literature-derived preference pairs.

3.2 Stage 2: Literature Preference Learning

Building on the resources constructed by EGM, LPL adapts the generator through two sequential stages: supervised fine-tuning for literature-derived synthesis and reasoning supervision, followed by direct preference optimization using experimentally supported preferences.

Stage 2a: Domain Knowledge Injection (SFT). We fine-tune Qwen3-8B (Yang et al., 2025) with LoRA (Hu et al., 2022; Ding et al., 2023) using three EGM-derived generation tasks: target-conditioned synthesis-plan generation, baseline-to-novel catalyst transformation, and evidence-grounded question–answer generation. CORE_OER contributes all three task types, whereas REAGENT_OER contributes target-conditioned generation and question–answer examples because it lacks matched baseline comparisons. Generation targets follow Qwen3's native `<think>` format to incorporate literature-derived reasoning supervision alongside structured synthesis output;

detailed example construction is provided in Appendix B. The best checkpoint is selected by validation token accuracy with early stopping.

Stage 2b: Literature Preference Alignment (DPO). We apply DPO (Rafailov et al., 2023) to CORE_OER preference pairs under the corresponding performance context:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(\mathbf{x},y_w,y_l)\sim\mathcal{D}_{\text{DPO}}} \log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid \mathbf{x})}{\pi_{\text{ref}}(y_w \mid \mathbf{x})} - \beta \log \frac{\pi_\theta(y_l \mid \mathbf{x})}{\pi_{\text{ref}}(y_l \mid \mathbf{x})}\right). \tag{1}$$

Here, y_w and y_l correspond to the reported better-performing and baseline synthesis plans, respectively, π_{ref} is the SFT checkpoint, and β=0.1 controls deviation from the reference policy. Chosen and rejected outputs use the same <think> format skeleton to preserve the structured behavior learned during SFT and focus preference learning on substantive synthesis differences. The best checkpoint is selected using `eval_rewards/margins`.

3.3 Stage 3: Agentic Closed-Loop Optimization

Building on the preference-aligned generator and the EGM retrieval knowledge base, the agentic optimizer performs inference-time synthesis-plan refinement through an evaluate–retrieve–revise loop. The generator acts as a search operator over executable synthesis plans, guided by domain-specific evaluation feedback and retrieved literature evidence. We maintain evaluator–generator independence to reduce self-referential bias, as using the same model for generation and evaluation may favor linguistic plausibility over scientific validity (Zheng et al., 2023; Wataoka et al., 2025).

Optimization objective. Given an initial synthesis plan $\text{Plan}^{(0)}$, the optimizer generates a sequence of revised plans:

$$\text{Plan}^{(t+1)} = G\Big(\text{Plan}^{(t)}, \mathcal{E}(\text{Plan}^{(t)}), \mathcal{R}\Big(\mathcal{E}(\text{Plan}^{(t)})\Big)\Big), \tag{2}$$

where $\mathcal{E}$ denotes the evaluator, $\mathcal{R}$ the literature retriever, and G the generator. The loop terminates when the overall evaluator score exceeds τ=0.9, when no improvement is observed for three consecutive rounds, or after a maximum of 12 rounds.

Hybrid Evaluator. The evaluator assesses each synthesis plan across nine dimensions using LLM-based judgments for domain-sensitive criteria and rule-based checks for verifiable constraints (Appendix C). Calibrated against domain-expert assessments and experimentally informed criteria, it provides both an overall optimization score and diagnostic feedback identifying deficiencies in the current plan. The evaluator serves as a surrogate signal for iterative refinement and early screening rather than a replacement for physical validation.

Agentic Retriever. The retriever grounds each refinement step in the EGM knowledge base. At each iteration, evaluator feedback identifies deficiencies and is converted into targeted retrieval queries. The retriever then returns relevant and diverse literature-grounded synthesis examples that address these deficiencies and provide evidence for subsequent revision. Unlike static retrieval augmentation, the retrieved context is updated as the weaknesses of the current plan evolve, providing iteration-specific evidence for subsequent revision.

Generator. The generator G acts as the search operator in the evaluate–retrieve–revise loop, integrating the current plan $\text{Plan}^{(t)}$, evaluator feedback $(s^{(t)}, \delta^{(t)})$, and retrieved evidence $\mathcal{R}^{(t)}$ to produce $\text{Plan}^{(t+1)}$. Rather than selecting from a predefined candidate set, it directly revises executable synthesis plans in response to iteration-specific feedback and evidence. Generation uses temperature 0.8 to encourage exploration of alternative synthesis strategies.

4 EXPERIMENTS

We evaluate ONEMAT across four complementary aspects: evidence-grounded mining, synthesis-plan generation, closed-loop optimization, and prospective wet-lab validation.

4.1 EXPERIMENTAL SETUP

Training setup. We fine-tune Qwen3-8B (Yang et al., 2025) with LoRA (Hu et al., 2022; Ding et al., 2023) (r=64, α=128). SFT is performed for 3 epochs with a learning rate of 2×10^{-5} and cosine decay, followed by 3 epochs of DPO (Rafailov et al., 2023) with a learning rate of 5×10^{-6} and β=0.1. Both stages use a maximum sequence length of 2,048, bf16 precision, and eight NVIDIA RTX 4090 GPUs. We compare **Base**, **SFT w/o CoT**, **SFT**, and **DPO**; SFT w/o CoT uses the same data and optimization settings as SFT but removes the <think> reasoning blocks.

Evaluation protocol. Mining quality is evaluated on 100 randomly sampled papers using blind assessment by three annotators together with predefined extraction-quality metrics. Synthesis-plan generation is evaluated on a 180-paper held-out set excluded from both training resources and the retrieval knowledge base. Closed-loop optimization is evaluated across generator variants under the same evaluate–retrieve–revise framework.

4.2 MINING QUALITY EVALUATION

To evaluate the reliability of EGM for constructing evidence-grounded scientific knowledge from literature, we perform a 4×3 ablation study across four mining models (Qwen3.7-Max, Gemini-3.1-Pro, GPT-5.5, and DeepSeek-V4-Pro) and three evidence-ordering configurations. Configuration A performs extraction without explicit evidence fields, configuration B generates extracted attributes before source evidence, and configuration C follows the proposed evidence-first ordering, where `source_evidence` is generated before extracted attributes.

We evaluate mining quality using five complementary metrics: **Field Accuracy** (FA), expert-rated on a 1–5 scale and normalized to 0–1; **Hallucination Rate** (HR), defined as the fraction of unsupported extracted values based on expert assessment; **Field Recall** (FR), measuring the fraction of recoverable fields extracted from the source document; **Evidence Consistency** (EC), calculated from the semantic similarity between `source_evidence` and extracted attributes for configurations with explicit evidence fields; and **Format Compliance** (FC), measuring validity against the predefined JSON schema.

Figure 4 summarizes the ablation results. Across all evaluated LLMs, the evidence-first configuration (C) consistently improves extraction quality, achieving higher FA, FR, EC, and FC while reducing HR compared with both no-evidence (A) and evidence-last (B) settings. These results demonstrate the effectiveness of evidence-first extraction for constructing reliable synthesis knowledge from scientific literature. Importantly, the comparison between evidence-last (B) and evidence-first (C) isolates the effect of evidence ordering, as both settings provide access to supporting evidence. The consistent improvement of C over B demonstrates that evidence availability alone is insufficient; instead, evidence must participate in the generation process to guide extraction decisions. The robustness of this improvement across four LLM backbones further indicates that the benefit arises from the evidence-grounded extraction strategy rather than a particular model backbone.

4.3 SYNTHESIS-PLAN GENERATION EVALUATION

Single-shot proposal quality. Figure 1(a) compares four generator variants, **Base**, **SFT w/o CoT**, **SFT**, and **DPO**, where DPO is applied on top of the SFT checkpoint, across eight generator-level evaluation dimensions on the 180-paper held-out test set. This comparison evaluates how literature-derived supervision affects single-shot synthesis-plan generation under controlled post-training settings. All models are conditioned on identical performance targets. Uniqueness is excluded from this analysis because its retrieval-based score depends on the external knowledge base and therefore does not exclusively reflect generator capability.

SFT w/o CoT improves over the Base model across all eight dimensions and raises the overall evaluator score from 0.51 to 0.57, with the largest gains in Completeness (0.52→0.73) and Justification

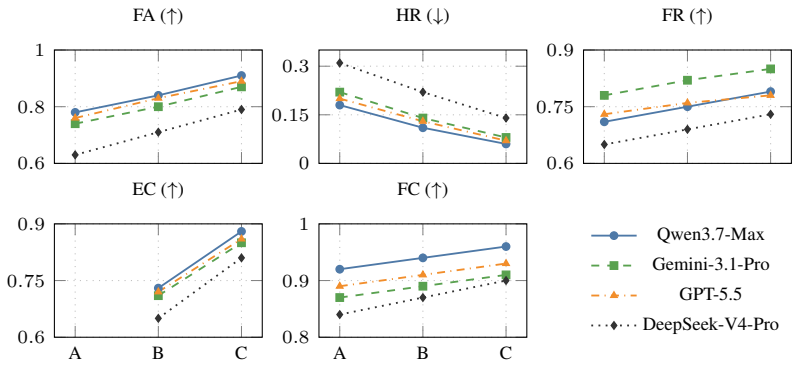


Figure 4: Evaluation of evidence-grounded mining quality under different models and evidence-ordering configurations. Four mining models are compared under three configurations: A (no evidence), B (evidence-last), and C (evidence-first). Higher values indicate better performance for FA, FR, EC, and FC, whereas lower values indicate better performance for HR. EC is only defined for configurations with explicit evidence fields.

(0.43→0.58). This improvement indicates that literature-derived synthesis examples provide effective domain supervision beyond the pretrained model's prior knowledge. Adding EGM-derived reasoning supervision further increases the overall score to 0.62 and improves all eight dimensions, including Completeness (0.73→0.88) and Justification (0.58→0.72), indicating that reasoning supervision provides additional benefits beyond domain adaptation alone. DPO further raises the score to 0.69 and improves six of the eight dimensions, including Stability, Activity, Novelty, and Cost. Completeness and Justification decrease only slightly (0.03 and 0.02), suggesting that preference alignment improves performance-oriented proposal quality while largely retaining the structured generation characteristics learned during SFT.

4.4 Closed-loop Optimization Evaluation

Iterative refinement behavior. Figure 1(b) compares the four generator variants under the evaluate–retrieve–revise loop. Across Base, SFT w/o CoT, and SFT, the fraction of runs whose peak score exceeds the initial score rises from 32% to 47% and 62%, with mean peak gains of $+0.04$, $+0.08$, and $+0.10$. However, these gains are not consistently retained at the endpoint: only 22%, 25%, and 30% of runs finish above their initial score, with mean final changes of -0.09, -0.08, and -0.07, while the fraction finishing within 0.05 of the within-run peak falls from 30% to 22% and 14%, indicating that intermediate improvements are frequently lost during subsequent refinement.

DPO improves both discovery and retention rather than merely increasing single-shot proposal quality: 94% of runs reach a peak above the initial score with a mean peak gain of $+0.15$, while 71% finish above the initial score, the mean final change reaches $+0.10$, and 63% finish within 0.05 of the within-run peak. These results indicate that supervised adaptation mainly improves the discovery of higher-scoring intermediate plans, whereas literature-derived preference alignment extends these gains to retention and endpoint performance.

Representative optimization trajectory. Figure 5 shows a representative DPO optimization trajectory for the 5Ir-Co_3O_4-bilayer design. The evaluator score increases from 0.48 to 0.74 after the first revision and to 0.84 after the second, reaches a peak of 0.86 at round 8, and finishes at 0.81 after 12 rounds, corresponding to a peak gain of $+0.38$ and a final gain of $+0.33$.

The trajectory is non-monotonic, reflecting a trade-off between performance and cost objectives. Higher Ir loading may improve catalytic performance, but the observed $\sim$33 wt% Ir content exceeds the target of $\leq$5 wt%. Reducing Ir usage can improve cost-related criteria while degrading performance-related scores, whereas subsequent revisions that recover performance may again favor higher Ir loading. This trade-off illustrates the difficulty of balancing catalytic performance against noble-metal utilization and explains why the best intermediate plan can differ from the final plan.

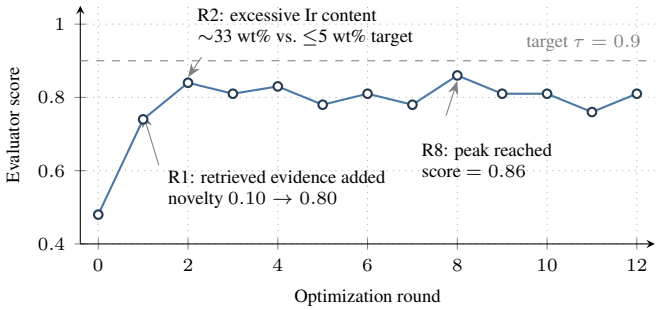


Figure 5: Representative DPO optimization trajectory for the 5Ir-Co_3O_4-bilayer design. The evaluator score increases from 0.48 initially to a within-run peak of 0.86 at round 8 and finishes at 0.81 after 12 rounds. The non-monotonic trajectory reflects competing optimization objectives, including catalytic performance and Ir utilization.

4.5 Experimental Validation

To evaluate whether OneMat-generated synthesis plans can be translated into physical experiments, we perform prospective wet-lab validation using an automated experimental workflow. The generated composition and synthesis strategy are directly executed without manual redesign. Figure 6 in Appendix D summarizes the automated workflow.

Among the experimentally executed designs, the best-performing catalyst is a RuO_2/MnO_2 heterostructure. In acidic OER evaluation, it achieves an overpotential of 155 mV at 10 mA cm^{-2} and exhibits an overpotential increase of only 21 mV after 500 CV cycles, demonstrating both high catalytic activity and electrochemical stability. These results provide strong experimental evidence that OneMat-generated synthesis plans can be translated into physically realizable, high-performing catalysts. The corresponding catalyst formulation, synthesis procedure, electrochemical evaluation, and characterization settings are reported in Appendix D.

5 Conclusion

We presented OneMat, a framework for synthesis-oriented generative optimization from scientific literature. OneMat transforms heterogeneous literature into an evidence-grounded domain resource through EGM, adapts an LLM-based synthesis-plan generator using literature-derived generation, reasoning, and preference supervision, and performs iterative refinement through an evaluate–retrieve–revise loop in which the generator acts as a search operator over executable synthesis plans. On acidic OER catalyst design, evidence-first extraction consistently improves mining quality across multiple LLMs, literature-derived supervision improves single-shot synthesis-plan generation, and preference alignment further improves both the discovery and retention of higher-scoring plans during closed-loop optimization. Prospective wet-lab validation further shows that generated synthesis plans can be translated into physically realizable, high-performing catalysts, with the best validated RuO_2/MnO_2 design achieving a low overpotential of 155 mV at 10 mA cm^{-2} and an increase of only 21 mV after 500 CV cycles. Together, these results support scientific literature as both a knowledge source and a supervision substrate for generative optimization toward experimentally realizable and high-performance materials synthesis.

Limitations and future work. The current framework relies on a Hybrid Evaluator as a surrogate for experimental performance. Although informed by domain-expert assessment and wet-lab outcomes, the evaluator cannot fully capture trade-offs among competing objectives or replace physical validation. Evidence-first mining is evaluated across four LLMs but only within acidic OER. The optimization loop remains evaluator-guided, with wet-lab measurements used for validation and calibration rather than online feedback. Future work will close this model–experiment loop by incorporating experimental outcomes into model training and inference-time optimization. Further directions include extending OneMat to broader materials and chemistry domains and improving multi-objective optimization across catalytic performance, cost, and resource utilization.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting portions of the manuscript, translation, language editing and rewriting, as well as code development and debugging. We did not use generative AI tools to develop the conceptual or theoretical framework, formulate research hypotheses, design the research methodology or experiments, clean or reformat datasets, or interpret experimental results. Other disclosure categories are not applicable to this work. All AI-assisted content was reviewed by the authors. Reported values were cross-checked against the underlying data, experiment logs, and configuration files, and citations and source materials were manually verified. The LLMs used as components of the proposed method, including literature-grounded structured extraction, are described in Section 3.1 and Appendix B and are distinct from the generative AI assistance disclosed above. We take full responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological details, training settings, evaluation protocols, and ablation configurations required to reproduce the computational experiments in the main paper and appendices. The construction of the literature-derived dataset, retrieval knowledge base, supervised examples, and preference pairs is described in Section 3.1, while generator adaptation and closed-loop optimization settings are detailed in Sections 3.2 and 3.3. Experimental evaluation protocols and additional analyses are reported in Section 4. For wet-lab validation, Appendix D summarizes the automated experimental workflow and reports the available catalyst formulation, synthesis, electrochemical, and characterization settings. All code, dataset, and checkpoints will be released on GitHub and Hugging Face upon acceptance.

REFERENCES

Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross King. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of The Royal Society Interface*, 22(227):20240674, June 2025. doi: 10.1098/rsif.2024.0674. URL https://doi.org/10.1098/rsif.2024.0674.

Mehrad Ansari and Seyed Mohamad Moosavi. Agent-based learning of materials datasets from the scientific literature. *Digital Discovery*, 3(12):2607–2617, December 2024. doi: 10.1039/d4dd00252k. URL https://doi.org/10.1039/d4dd00252k.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022. URL https://arxiv.org/abs/2212.08073v1.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelli-*

gence, 6(5):525–535, May 2024. doi: 10.1038/s42256-024-00832-8. URL https://doi.org/10.1038/s42256-024-00832-8.

Andres M. Bran, Théo A. Neukomm, Daniel Armstrong, Zlatko Jončev, and Philippe Schwaller. Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation. *Matter*, 9(5):102812, May 2026. doi: 10.1016/j.matt.2026.102812. URL https://doi.org/10.1016/j.matt.2026.102812.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583 (7815):237–241, July 2020. doi: 10.1038/s41586-020-2442-2. URL https://doi.org/10.1038/s41586-020-2442-2.

Honghao Chen, Hongxuan Liu, Yishen Tew, Xiaotian Ren, Xiaojin Tang, and Xiaonan Wang. Distilling knowledge from catalysis literature with long-context large language model agents. *ACS Catalysis*, 15(21):18244–18254, November 2025. doi: 10.1021/acscatal.5c06431. URL https://doi.org/10.1021/acscatal.5c06431.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, March 2023. doi: 10.1038/s42256-023-00626-4. URL https://doi.org/10.1038/s42256-023-00626-4.

Simon Geiger, Olga Kasian, Marc Ledendecker, Enrico Pizzutilo, Andrea M. Mingers, Wen Tian Fu, Oscar Diaz-Morales, Zhizhong Li, Tobias Oellers, Luc Fruchter, Alfred Ludwig, Karl J. J. Mayrhofer, Marc T. M. Koper, and Serhiy Cherevko. The stability number as a metric for electrocatalyst stability benchmarking. *Nature Catalysis*, 1(7):508–515, July 2018. doi: 10.1038/s41929-018-0085-6. URL https://doi.org/10.1038/s41929-018-0085-6.

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: 10.1002/adma.202413523. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202413523.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xu Hu, Suya Chen, Letian Chen, Huijuan Wang, Xu Zhang, and Zhen Zhou. Automating structure–activity analysis for electrochemical nitrogen reduction catalyst design through multi-agent collaborations. *National Science Review*, 12(11):nwaf372, November 2025. doi: 10.1093/nsr/nwaf372. URL https://doi.org/10.1093/nsr/nwaf372.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, February 2024. doi: 10.1038/s42256-023-00788-1. URL https://doi.org/10.1038/s42256-023-00788-1.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, March 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17464–17497. Curran Associates, Inc., 2023. doi: 10.52202/075280-0767. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/38b787fc530d0b31825827e2cc306656-Paper-Conference.pdf.

Jakob Kibsgaard and Ib Chorkendorff. Considerations for the scaling-up of water splitting catalysts. *Nature Energy*, 4(6):430–433, June 2019. doi: 10.1038/s41560-019-0407-1. URL https://doi.org/10.1038/s41560-019-0407-1.

Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, November 2017. doi: 10.1021/acs.chemmater.7b03500. URL https://doi.org/10.1021/acs.chemmater.7b03500.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203, October 2019. doi: 10.1038/s41597-019-0224-1. URL https://doi.org/10.1038/s41597-019-0224-1.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024. URL https://openreview.net/forum?id=AAxIs3D2ZZ.

Zhanwu Lei, Yan Huang, Yuanmin Zhu, Donglai Zhou, Yu Chen, Song Wang, Wanxia Li, Huirong Li, Xiaoke Xi, Yang Liu, Yuchen Zhang, Guozhen Zhang, Xiyu Li, Qing Zhu, Baicheng Zhang, Shuo Feng, Sheng Ye, Wensheng Yan, Shuo Zhang, Shuhong Jiao, Jun Jiang, Meng Gu, Ruiguo Cao, and Yi Luo. Automatic discovery and optimal generation of amorphous high-entropy electrocatalysts. *Journal of the American Chemical Society*, 147(25):21743–21753, June 2025. doi: 10.1021/jacs.5c04117. URL https://doi.org/10.1021/jacs.5c04117.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Turab Lookman, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, February 2019. doi: 10.1038/s41524-019-0153-8. URL https://doi.org/10.1038/s41524-019-0153-8.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/

075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Charles C. L. McCrory, Suho Jung, Jonas C. Peters, and Thomas F. Jaramillo. Benchmarking heterogeneous electrocatalysts for the oxygen evolution reaction. *Journal of the American Chemical Society*, 135(45):16977–16987, November 2013. doi: 10.1021/ja407115p. URL https://doi.org/10.1021/ja407115p.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, December 2023. doi: 10.1038/s41586-023-06735-9. URL https://doi.org/10.1038/s41586-023-06735-9.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating materials with riemannian flow matching. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35664–35686. PMLR, July 2024. URL https://proceedings.mlr.press/v235/miller24a.html.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek (eds.), *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 56–64, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4007. URL https://aclanthology.org/W19-4007/.

Hong Nhan Nong, Lorenz J. Falling, Arno Bergmann, Malte Klingenhof, Hoang Phi Tran, Camillo Spöri, Rik Mom, Janis Timoshenko, Guido Zichittella, Axel Knop-Gericke, Simone Piccinin, Javier Pérez-Ramírez, Beatriz Roldan Cuenya, Robert Schlögl, Peter Strasser, Detre Teschner, and Travis E. Jones. Key role of chemistry versus bias in electrocatalytic oxygen evolution. *Nature*, 587(7834):408–413, November 2020. doi: 10.1038/s41586-020-2908-2. URL https://doi.org/10.1038/s41586-020-2908-2.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko,

M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Andrei Rekesh, Miruna Cretu, Dmytro Shevchuk, Pietro Lio, Robert A. Batey, Mike Tyers, Michał Koziarski, and Cheng-Hao Liu. SynCoGen: Synthesizable 3D molecule generation via joint reaction and coordinate modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=24QKU4iqft.

Isolda Roger, Michael A. Shipman, and Mark D. Symes. Earth-abundant catalysts for electrochemical and photoelectrochemical water splitting. *Nature Reviews Chemistry*, 1(1):0003, January 2017. doi: 10.1038/s41570-016-0003. URL https://doi.org/10.1038/s41570-016-0003.

Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 54(3):1125–1150, February 2025. doi: 10.1039/d4cs00913d. URL https://doi.org/10.1039/d4cs00913d.

Zhi Wei Seh, Jakob Kibsgaard, Colin F. Dickens, Ib Chorkendorff, Jens K. Nørskov, and Thomas F. Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321):eaad4998, 2017. doi: 10.1126/science.aad4998. URL https://www.science.org/doi/abs/10.1126/science.aad4998.

Linsey C. Seitz, Colin F. Dickens, Kazunori Nishio, Yasuyuki Hikita, Joseph Montoya, Andrew Doyle, Charlotte Kirk, Aleksandra Vojvodic, Harold Y. Hwang, Jens K. Nørskov, and Thomas F. Jaramillo. A highly active and stable $IrO_x/SrIrO_3$ catalyst for the oxygen evolution reaction. *Science*, 353(6303):1011–1014, 2016. doi: 10.1126/science.aaf5050. URL https://www.science.org/doi/abs/10.1126/science.aaf5050.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218. URL https://doi.org/10.1109/JPROC.2015.2494218.

Matthew C. Swain and Jacqueline M. Cole. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, October 2016. doi: 10.1021/acs.jcim.6b00207. URL https://doi.org/10.1021/acs.jcim.6b00207.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, December 2023. doi: 10.1038/s41586-023-06734-w. URL https://doi.org/10.1038/s41586-023-06734-w.

Daniel P. Tabor, Loïc M. Roch, Semion K. Saikin, Christoph Kreisbeck, Dennis Sheberla, Joseph H. Montoya, Shyam Dwaraknath, Muratahan Aykol, Carlos Ortiz, Hermann Tribukait, Carlos Amador-Bedolla, Christoph J. Brabec, Benji Maruyama, Kristin A. Persson, and Alán Aspuru-Guzik. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials*, 3(5):5–20, May 2018. doi: 10.1038/s41578-018-0005-z. URL https://doi.org/10.1038/s41578-018-0005-z.

Ludi Wang, Xueqing Chen, Yi Du, Yuanchun Zhou, Yang Gao, and Wenjuan Cui. CataLM: empowering catalyst design through large language models. *International Journal of Machine Learning and Cybernetics*, 16(5):3681–3691, June 2025. doi: 10.1007/s13042-024-02473-0. URL https://doi.org/10.1007/s13042-024-02473-0.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge, 2025. URL https://openreview.net/forum?id=Ns8zGZ0lmM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, July 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, March 2025. doi: 10.1038/s41586-025-08628-5. URL https://doi.org/10.1038/s41586-025-08628-5.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Peng Zheng, Zhennan Han, Bao-Lian Su, and Guangwen Xu. Catal-GPT: AI-driven directed efficient design framework for catalysts. *National Science Review*, 12(9):nwaf299, September 2025. doi: 10.1093/nsr/nwaf299. URL https://doi.org/10.1093/nsr/nwaf299.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart (eds.), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://aclanthology.org/2024.nlp4science-1.10/.

A EGM Output Schema

EGM uses two closely related JSON schemas corresponding to the comparative and synthesis-focused literature records introduced in Section 3.1. Comparative records use a paired schema containing `source_evidence`, `novel_catalyst`, `baseline_catalyst`, `explanation`, and

cots. The novel_catalyst and baseline_catalyst fields share the same internal structure, including catalyst identity, synthesis details, and evaluation metrics. In both schema variants, source_evidence is generated before the extracted attributes, implementing the evidence-first ordering described in Section 3.1.

Comparative-record schema. The comparative schema is summarized below, with repeated substructures and lower-level metric fields abbreviated for compactness; baseline_catalyst follows the same internal schema as novel_catalyst.

```
{
    "source_evidence": {
        "synthetic_method": {"locator": null, "raw": null},
        "reagents": {"locator": null, "raw": null},
        "activity_metrics": {"locator": null, "raw": null},
        "stability_metrics": {"locator": null, "raw": null},
        ...
    },

    "novel_catalyst": {
        "catalyst_name": null,
        "synthesis_details": {
            "synthetic_method": null,
            "reagents": [
                {
                    "types": null,
                    "species": null,
                    "amount": null
                }
            ],
            "preparation": {
                "temperature": null,
                "time": null,
                "pH": null,
                "procedures": null
            },
            "purification_before_calcination": null,
            "calcination": {
                "temperature": null,
                "time": null,
                "gas": null,
                "procedures": null
            },
            "purification_after_calcination": null
        },
        "evaluation_metrics": {
            "stability": {
                "operation_time": {...},
                "overpotential_increase": {...},
                "degradation_rate": {...},
                "accelerated_stress_tests": {...}
            },
            "activity": {
                "overpotential_at_10_mA_cm2": null,
                "overpotential_at_100_mA_cm2": null
            },
            "element_content": {
                "ir": null,
                "ru": null
            }
```

```
        }
    },

    "baseline_catalyst": {
        "...": "same schema as novel_catalyst"
    },

    "explanation": null,
    "cots": [
        {
            "query": null,
            "cot": null,
            "answer": null
        }
    ]
}
```

The source_evidence object stores both a source locator and the corresponding verbatim text for each supported attribute. The synthesis sub-schema records the synthetic method, reagents, preparation conditions, purification steps, and calcination conditions, while the evaluation sub-schema captures stability, activity, and elemental-content information. The explanation field records the reported design rationale, and cots contains reasoning-oriented question–answer entries derived from the source context.

Synthesis-focused records. For records without a matched baseline catalyst, EGM uses a simplified single-catalyst schema. It retains source_evidence, explanation, and cots, while replacing the paired novel_catalyst/baseline_catalyst structure with the top-level fields catalyst_name, synthesis_details, and evaluation_metrics. The internal synthesis and evaluation sub-schemas are identical to those used for each catalyst in the comparative schema.

B SFT Example Construction

The structured records produced by EGM are transformed into three complementary types of supervised fine-tuning examples used for generator adaptation in Section 3.2. For synthesis-plan generation examples, the reasoning target is constructed directly from the mined record. Specifically, the record-level explanation is first used as an overall design rationale, followed by the entries in cots, each formatted as a question, intermediate reasoning trace, and answer. These components are concatenated into a single reasoning chain enclosed by Qwen3's <think> and </think> tags, followed by the structured synthesis_details output.

- **Target-conditioned synthesis-plan generation.** The input contains the reported evaluation metrics as reference_metrics together with available activity and stability evidence extracted from source_evidence. The target consists of the constructed reasoning block followed by the corresponding structured synthesis plan. Both CORE_OER and REAGENT_OER contribute examples of this type.
- **Baseline-to-novel catalyst transformation.** For comparative CORE_OER records with an available baseline, the input contains the baseline synthesis plan and its reported evaluation metrics. The target uses the same reasoning-plus-synthesis structure as the target-conditioned examples and corresponds to the literature-reported novel catalyst. These examples provide optimization-oriented supervision for transforming an existing synthesis plan into an experimentally reported improved alternative.
- **Evidence-grounded reasoning Q&A.** Each complete cots entry is additionally expanded into an individual supervised example. The extracted query forms the input, while the output contains the corresponding cot enclosed in a <think> block followed by the extracted answer. These examples provide direct reasoning supervision grounded in the source-derived scientific context.

CORE_OER contributes all three example types, whereas REAGENT_OER contributes target-conditioned synthesis-plan generation and evidence-grounded reasoning Q&A because it does not contain matched baseline comparisons. Together, these transformations produce 29,148 supervised fine-tuning examples used in the SFT stage of LPL.

C HYBRID EVALUATOR DETAILS

The Hybrid Evaluator combines LLM-based scientific assessment with rule-based checks for verifiable constraints. Table 1 summarizes the nine evaluation dimensions, their contributions to the overall score, and the corresponding assessment mechanisms. Stability Potential, Activity Potential, Novelty, and Cost Rationality receive the largest weights, while Plausibility, Justification, and Field Completeness provide complementary assessments. Feasibility acts as a hard veto, whereas Uniqueness is used as an auxiliary retrieval-based diagnostic and is not included in the weighted overall score.

Table 1: Hybrid Evaluator dimensions and assessment methods. *Core dimensions receive the largest weights in the overall evaluator score. Feasibility acts as a hard veto, while Uniqueness is reported separately from the weighted score.

Dimension	**Weight**	**Method**	**Key criterion**
Stability Potential*	30%	LLM-based	Stability under OER conditions
Activity Potential*	30%	LLM-based	Activity toward target metrics
Novelty*	12.5%	LLM-based	Distinct composition or synthesis strategy
Cost Rationality*	12.5%	LLM-based	Ru $\leq$ 20 wt%, Ir $\leq$ 5 wt%
Plausibility	5%	Rule + reagent DB	Balance of known and unfamiliar reagents
Justification	5%	LLM-based	Literature-grounded design rationale
Field Completeness	5%	Schema validation	Required synthesis fields present
Feasibility	veto	Rule-based	Score $= 0$ if a hard constraint is violated
Uniqueness	–	Vector retrieval	Similarity to literature formulations

Evaluator calibration. Because the Hybrid Evaluator provides the optimization signal used throughout the agentic loop, its overall score is calibrated against independent domain-expert assessments. Experts assign holistic scores to held-out synthesis plans while remaining blind to both generator identity and evaluator scores, and agreement is measured using Spearman's rank correlation coefficient ρ. The relative weights of the evaluation dimensions are adjusted to improve agreement with expert assessments and are subsequently checked against available wet-lab outcomes to improve consistency with experimentally observed catalyst quality. The resulting evaluator configuration is fixed before the reported closed-loop optimization experiments.

Plausibility. Let r denote the fraction of proposed reagents not found in the known-reagent library. The Plausibility score is defined as

$$S_{\text{plaus}} = \max\big(0.2,\ 1.0 - 4(r - 0.25)^2\big)\,.$$

The score is maximized at $r = 0.25$, favoring synthesis plans that rely predominantly on established reagents while allowing a limited proportion of less familiar components.

Uniqueness. Let $x \in [0, 1]$ denote the cosine similarity between a proposed synthesis plan and its most similar formulation in the vector knowledge base, and let $k \geq 0$ denote the strictness parameter. The Uniqueness score is defined as

$$U(x, k) = \begin{cases} 1 - x, & k = 0, \\ \dfrac{e^{-kx} - e^{-k}}{1 - e^{-k}}, & k > 0. \end{cases}$$

Larger values of k impose a steeper penalty on highly similar formulations. Uniqueness is used to characterize how distinct a proposal is from the literature knowledge base and is reported separately from the weighted overall evaluator score.

D AUTOMATED WET-LAB VALIDATION

Figure 6 summarizes the automated experimental workflow used for prospective validation of ONEMAT-generated synthesis plans. The platform integrates automated powder weighing and liquid dispensing, liquid-phase synthesis, positive-pressure solid–liquid separation, post-synthesis annealing, structural characterization by XRD and SEM, and electrochemical evaluation. In the experiments reported here, wet-lab measurements are used for experimental validation and evaluator calibration rather than as online feedback to the optimization loop.

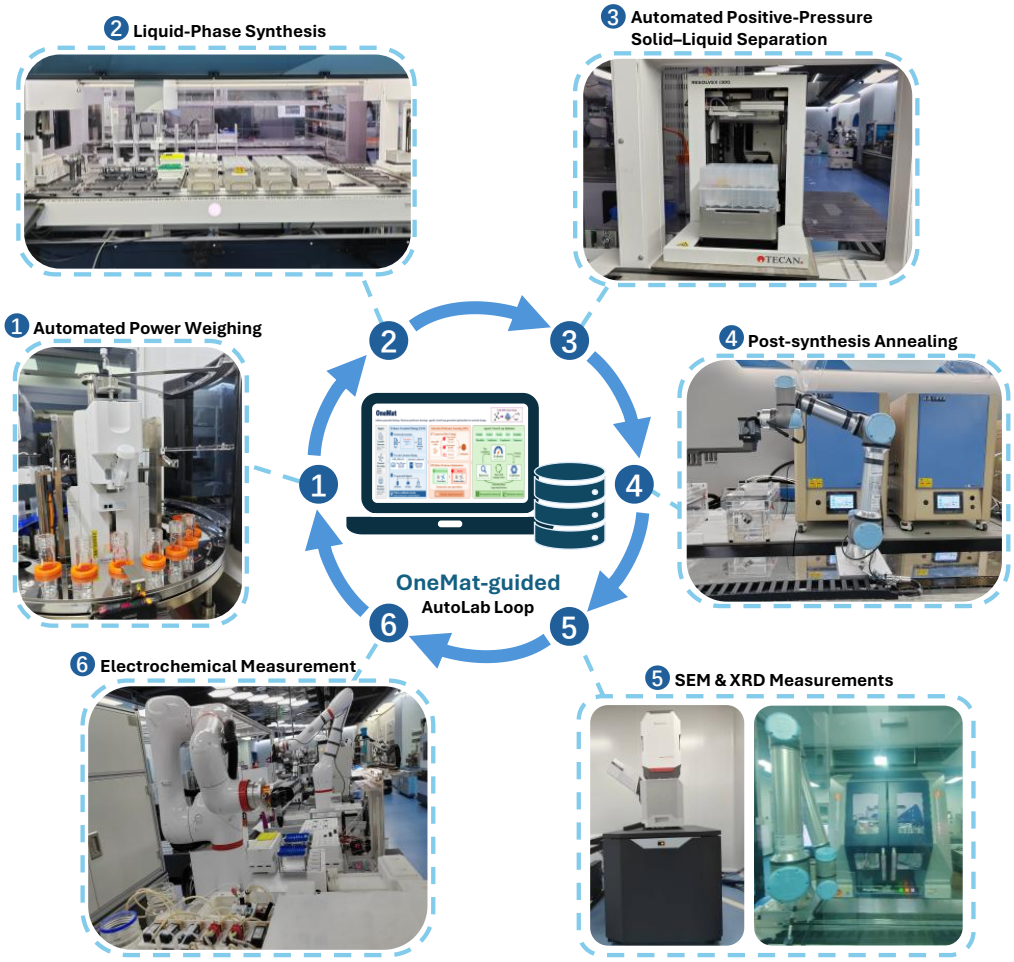


Figure 6: OneMat-guided AutoLab workflow for automated experimental validation. The platform integrates automated powder weighing and liquid dispensing, liquid-phase synthesis, positive-pressure solid–liquid separation, post-synthesis annealing, structural characterization by XRD and SEM, and electrochemical evaluation. Wet-lab measurements are used for experimental validation and evaluator calibration rather than as online feedback to the optimization loop.

Best-performing validated synthesis plan. The best-performing experimentally validated design reported in Section 4.5 is a RuO_2/MnO_2 heterostructure with a Ru:Mn elemental ratio of 3:2. The synthesis plan uses $RuCl_3$ (0.10 g, supplied as 10 mL of a 0.01 g mL^{-1} solution), $MnSO_4{\cdot}5H_2O$ (0.43 g), $(NH_4)_2S_2O_8$ (0.57 g), and deionized water (10 mL). $MnSO_4{\cdot}5H_2O$ and $(NH_4)_2S_2O_8$ were dissolved in deionized water, followed by addition of the $RuCl_3$ solution. The mixture was stirred for 10 min, transferred to an autoclave, and heated at 140 °C for 12 h. After cooling, the resulting black precipitate was collected, washed with deionized water and ethanol, and dried overnight at 60 °C. The dried product was subsequently annealed in air at 350 °C for 4 h with a

heating rate of 10 °C min^{-1}, followed by washing with deionized water. The generated plan further hypothesizes that the RuO_2/MnO_2 heterostructure may modify the local electronic environment and interfacial properties, potentially improving acidic OER activity and stability; this rationale is treated as a design hypothesis rather than direct mechanistic evidence.

Electrochemical evaluation. Electrochemical measurements were performed on a CHI760E workstation in 0.5 M H_2SO_4 using a standard three-electrode configuration, with a Pt sheet as the counter electrode and Ag/AgCl (saturated KCl) as the reference electrode. The working-electrode ink was prepared by dispersing 5 mg of catalyst in 980 μL of ethanol and 20 μL of Nafion solution, followed by ultrasonication for 30 min. A 200 μL aliquot of the resulting ink was deposited onto a 1×1 cm^2 carbon-paper substrate, corresponding to a catalyst loading of 1 mg cm^{-2}.

All measured potentials were converted to the reversible hydrogen electrode (RHE) scale according to

$$E_{\mathrm{RHE}} = E_{\mathrm{Ag/AgCl}} + 0.059\,\mathrm{pH} + 0.197.$$

Before activity evaluation, the working electrode was stabilized by 20 cyclic-voltammetry (CV) cycles between 1.0 and 1.8 V versus RHE at a scan rate of 100 mV s^{-1}. Linear sweep voltammetry (LSV) was subsequently performed at 10 mV s^{-1} with 90% iR correction. The overpotential at 10 mA cm^{-2} was extracted from the polarization curve as the activity metric, following standard OER benchmarking practice (McCrory et al., 2013).

Catalyst stability was evaluated using an accelerated stress test consisting of 500 CV cycles between 1.0 and 1.8 V versus RHE at a scan rate of 50 mV s^{-1}. The overpotential at 10 mA cm^{-2} was measured again after cycling, and the corresponding change in overpotential, $\Delta\eta_{\mathrm{OER}}$, was used as the stability metric. The best-performing RuO_2/MnO_2 catalyst achieves an initial overpotential of 155 mV at 10 mA cm^{-2}, with an increase of 21 mV after 500 CV cycles.

Material characterization. Phase and crystalline structures were characterized by X-ray diffraction (XRD) using a Rigaku SmartLab diffractometer with a Cu source and a scanning rate of 10° min^{-1}. Scanning electron microscopy (SEM) was performed using an Apreo ChemiSEM for morphological and compositional characterization of the synthesized samples.