Retro-Forge: A Multi-Step Pairwise Retrosynthesis Framework for Solid-State Materials Synthesis

Anonymous Authors[1]

Abstract

AI-driven materials discovery has made remarkable strides in generating stable, unique, and novel structures at unprecedented scale, yet a critical gap remains between generation and realization: without valid synthesis routes, proposed materials stay in silico. Existing approaches to precursor prediction (PP)—the first and most consequential step of material synthesis planning (MSP)—formulate it as a single-step problem, overlooking the well-established domain knowledge that solid-state reactions proceed pairwise, while relying on text-mined datasets known to suffer from extraction errors, chemical invalidity, and systematic compositional bias. We introduce Retro-Forge, a multi-step pairwise retrosynthesis framework that for the first time casts PP as a sequence of learnable pairwise reactions. Built on a chemically valid Pairwise Reaction Dataset (PRD, ∼6k entries) and expanded through synthetic data augmentation to address data scarcity, a single-step pairwise reactant prediction model is trained and composed recursively via tree search to produce complete synthesis routes. Retro-Forge matches state-of-the-art PP baselines under a contamination-free evaluation protocol, demonstrating that the multi-step pairwise formulation is both learnable and effective—and that further advances in pairwise reactant prediction will directly translate to improved synthesis route discovery.

1. Introduction

The application of artificial intelligence to inorganic materials discovery has grown rapidly. Generative and predictive models now propose candidate materials at unprecedented scale (Merchant et al., 2023; Zeni et al., 2025), yet bringing them into the laboratory remains the central bottleneck: without valid synthesis routes, even the most promising predicted materials cannot be realized experimentally (Szymanski et al., 2023b). Materials synthesis planning (MSP)—determining how to produce a target material from available starting materials—is intrinsically difficult, admitting a one-to-many mapping in which a single target may be reachable through multiple routes involving different precursors, intermediates, and processing conditions. Among the subtasks of MSP, precursor prediction (PP)—selecting which starting materials react to form a given target—is the first and most consequential decision: errors here propagate through every subsequent step of the synthesis plan.

Previous approaches to PP fall along two broadly distinct lines. Early work adopted retrieval-based and similarity-driven strategies over known reactions (Huo et al., 2022; He et al., 2023), thermodynamic optimization over computed formation energies (Aykol et al., 2021), and ranking-based approaches that embed targets and precursor sets in a shared latent space (Prein et al., 2025a). More recently, large language models fine-tuned on synthesis corpora have achieved strong performance on precursor prediction (Prein et al., 2025b; Song et al., 2025; Noh et al., 2026). What these approaches share is a common formulation: precursor prediction is treated as a single-shot mapping from a target composition to a complete precursor set, bypassing the sequential structure of how solid-state reactions actually proceed.

In solid-state inorganic synthesis, experiments have shown that multi-component transformations proceed through sequences of *pairwise* reactions between adjacent reactant phases (Malkowski et al., 2021): exactly two solid phases react at their mutual interface, driven by diffusion-limited kinetics, and the product of one pairwise step may serve as a reactant in the next. For example, the well-known material $YBa_2Cu_3O_{6+x}$ (YBCO) is synthesized through an initial pairwise reaction of BaO_2 and CuO forming $Ba_2Cu_3O_6$, which later reacts with Y_2O_3 to form the target material (Miura et al., 2021). This principle has been translated into algorithmic frameworks that rank candidate pairwise reactions by thermodynamic driving force and have achieved experimental discovery of novel inor-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

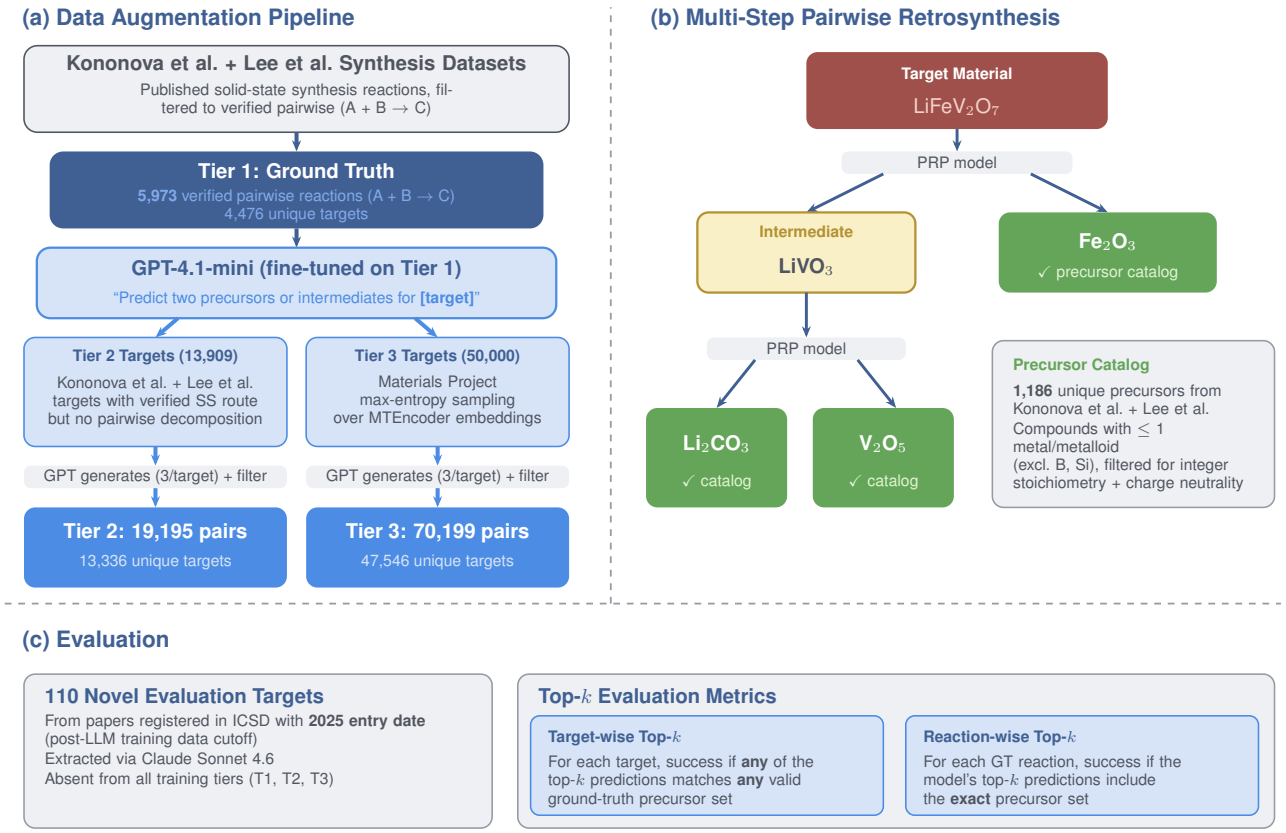

Figure 1. Overview of Retro-Forge. **(a)** Data augmentation pipeline. Verified pairwise reactions from the Kononova and Lee corpora (Kononova et al., 2019; Lee et al., 2025) form the Tier 1 ground truth (5,973 reactions). A GPT-4.1-mini model fine-tuned on Tier 1 generates 3 candidate pairwise reactant pairs for two additional target pools: Tier 2 (13,909 Kononova/Lee targets lacking pairwise annotations) and Tier 3 (50,000 Materials Project targets selected via maximum-entropy sampling). After filtering, the augmented corpus comprises 19,195 (Tier 2) and 70,199 (Tier 3) pairwise reactions. **(b)** Multi-step pairwise retrosynthesis. A single-step pairwise reactant prediction (PRP) model predicts two reactants for a given target. The process is applied recursively: intermediates (amber) are further decomposed until all leaf nodes are purchasable precursors found in the precursor catalog (green). The example shows $LiFeV_2O_7$ decomposed through the intermediate $LiVO_3$ into the final precursor set $\{Li_2CO_3,\ V_2O_5,\ Fe_2O_3\}$. **(c)** Evaluation. Models are evaluated on 110 contamination-free post-October-2024 targets—ensuring temporal separation from all training data—using a target-wise top-k metric, contrasted with the reaction-wise top-k metric used in prior work.

ganic phases in autonomous laboratory settings (Szymanski et al., 2023a;b). However, these approaches rely on rule-based heuristics computed from *ab initio* formation energies, which are computationally expensive and unavailable for the majority of candidate or intermediate materials. It remains an open question how to cast the pairwise reaction as a *learnable* unit trained directly on large-scale synthesis corpora—turning isolated recipes into recombinable building blocks, in the same spirit as organic retrosynthesis extends known reactions through learnable bond disconnections (Corey & Wipke, 1969).

Addressing this question also requires confronting two practical obstacles. First, the datasets underpinning existing models—most notably Kononova et al. (2019) and Lee et al. (2025)—contain substantial noise from automated extraction pipelines, including misassigned stoichiometries and conflated precursor and target species, with fewer than two-thirds of sampled entries found to be fully correct (Chung et al., 2025; Prein et al., 2025b). Beyond extraction errors, the corpora are systematically biased toward publisher-accessible literature, overrepresenting well-studied material families while leaving large regions of chemical space undercharacterized (Prein et al., 2025b; Sun & David, 2025). Second, for LLM-based approaches specifically, evaluation on targets predating the model pretraining cutoff makes it difficult to distinguish genuine generalization from knowledge recall (Magar & Schwartz, 2022; Sainz et al., 2023), calling for a contamination-free evaluation protocol.

In this paper, we introduce **Retro-Forge**, a multi-step pairwise retrosynthesis framework for solid-state inorganic synthesis that addresses the above obstacles. We construct a *Pairwise Reaction Dataset* (PRD), a chemically curated and validated dataset of solid-state pairwise reactions, by rigorously filtering the Kononova and Lee corpora (Kononova et al., 2019; Lee et al., 2025) to retain only verified, chemically valid pairwise reactions. We expand coverage through LLM-based data augmentation, in which a model fine-tuned on the PRD generates candidate pairwise reactions for a diverse pool of target compositions. A single-step pairwise reactant prediction model trained via staged fine-tuning on this augmented data is then deployed within a multi-step tree search that recursively decomposes a target into pairs of reactants until a complete synthesis route is formed. Critically, we evaluate Retro-Forge on targets drawn exclusively from papers published after October 2024—verified absent from all training data, including the pretraining corpora of the base LLMs—enabling a rigorous, contamination-free assessment of genuine generalization.

To the best of our knowledge, Retro-Forge is the first work to formulate precursor prediction for solid-state synthesis as a sequence of learnable pairwise reactions, and to evaluate model performance under a contamination-free protocol.

Our contributions are as follows:

- We propose Retro-Forge, a novel multi-step pairwise retrosynthesis framework that formulates PP as a sequence of learnable pairwise reactions, solved via a tree search over recursive two-reactant decompositions, enabling composition of pairwise reactions into complete synthesis pathways.
- We construct the first chemically curated pairwise reaction dataset for solid-state synthesis (PRD), comprising ∼6K chemically validated reactions.
- We propose a data augmentation pipeline in which an LLM fine-tuned on the PRD generates synthetic pairwise reactions from diverse target pools, demonstrating that augmentation substantially improves both single-step and multi-step synthesis planning performance.
- We conduct rigorous evaluation on 110 PP evaluation targets (post-October-2024) under a contamination-free protocol, showing that our pairwise approach achieves performance on par with the state-of-the-art single-step method with a straightforward and simple inference setup.

2. Related Works

Multi-step Retrosynthesis in Organic Chemistry. Retrosynthesis planning in organic chemistry, formalized by Corey and Wipke (Corey & Wipke, 1969) as the disconnection of a target molecule into simpler precursors via known reaction templates, has been extensively studied with machine learning. Modern approaches include template-based methods that retrieve and rank known reaction templates (Segler et al., 2018; Chen et al., 2020) and template-free methods that treat retrosynthesis as a sequence-to-sequence problem over molecular SMILES strings (Schwaller et al., 2019). Multi-step retrosynthesis extends single-step prediction with tree search algorithms—most notably MCTS and best-first search—to find complete routes from purchasable starting materials (Segler et al., 2018; Chen et al., 2020).

Despite the maturity of multi-step retrosynthesis in organic chemistry, analogous frameworks for inorganic solid-state synthesis remain absent. Existing text-mined datasets record synthesis reactions as one-shot recipes without capturing the underlying multi-step pairwise reaction structure, leaving this domain knowledge latent and unexploited—a gap that directly precludes data-driven multi-step retrosynthesis for solid-state synthesis planning.

Precursor Prediction. PP is a cornerstone of synthesis planning, tasked with identifying starting materials for target compounds. In the inorganic domain, this process has traditionally relied on density functional theory (DFT) to screen precursors based on thermodynamic stability (Engel & Dreizler, 2011). However, such calculations often fail to account for practical experimental preferences, such as kinetic accessibility (Szymanski & Bartel, 2024). To address these limitations, various materials encoding strategies have been proposed to represent inorganic compositions as numerical vectors, enabling similarity-based reasoning across materials (Ward et al., 2016; Wang et al., 2021). Such representations naturally lend themselves to retrieval-based frameworks, where precursor candidates are inferred by identifying synthesis precedents from similar target materials (He et al., 2023; Noh et al., 2025). Recently, elementwise representations of target compositions served as the inputs for the models (Kim et al., 2024; Prein et al., 2025a), enabling ranking of candidate precursor sets conditioned on target composition. While effective in improving precursor selection, these approaches still treat reactions as single-step events, without explicitly modeling intermediate transformations. This framing is particularly inadequate for solid-state synthesis, where stepwise pathways are required to capture the thermodynamic landscapes underlying distinct phase transformations.

LLMs for Materials Science. LLMs have emerged as a powerful paradigm for materials science by leveraging their broad knowledge of chemistry (Okabe et al., 2024). Recent works have explored their potential in precursor prediction from multiple perspectives. Prein et al. (2025b) utilized LLMs to generate synthetic reaction data to augment limited datasets, thereby boosting the accuracy of downstream expert models. This data-augmented approach complements

structural reasoning by allowing models to explore a broader chemical space and improving the robustness of condition predictions. In parallel, MSP-LLM (Noh et al., 2026) established a unified framework that organizes materials synthesis planning into a chemically consistent decision chain by predicting material group as an intermediate step. By explicitly conditioning the autoregressive decoding on hierarchical precursor types, such models have demonstrated that LLMs can go beyond simple text generation to designing coherent synthesis plans.

3. Preliminaries

Precursor Prediction (PP). The goal of PP is to predict a set of N precursor materials $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ required to synthesize a target material T, given only its chemical formula. No additional structural or contextual information is provided as input. For example, given the target $LiFeV_2O_7$, the task is to predict a precursor set such as $\{Li_2CO_3,\ V_2O_5,\ Fe_2O_3\}$. Since a single target material may be synthesized via multiple valid precursor combinations, evaluation is performed target-wise: a prediction is considered correct if the predicted set matches any one of the ground-truth precursor sets recorded for that target.

Pairwise Reaction. A pairwise reaction is a binary solid-state reaction in which exactly two solid reactants combine to form a single product. We denote a pairwise reaction as $\mathbf{p} = (p_1, p_2) \rightarrow t$, where p_1 and p_2 are the reactants and t is the product. For example, $(Li_2CO_3,\ V_2O_5) \rightarrow LiVO_3$ and $(LiVO_3,\ Fe_2O_3) \rightarrow LiFeV_2O_7$ are both pairwise reactions, where the left-hand side denotes the reactant pair $\mathbf{p}$ and the right-hand side denotes the product t.

Pairwise Reactant Prediction (PRP). Given a composition t, PRP is the reverse prediction problem of a pairwise reaction: predict a reactant pair $\mathbf{p} = (p_1, p_2)$ such that $p_1 + p_2 \rightarrow t$. Here t may be the final synthesis target T or an intermediate formed during the synthesis pathway, and likewise p_1 and p_2 may be purchasable starting materials or intermediates produced in earlier steps. For example, PRP of $LiFeV_2O_7$ as target t may result in $(LiVO_3,\ Fe_2O_3)$ as the reactant pair $\mathbf{p}$.

Pairwise Retrosynthesis. Pairwise retrosynthesis is a backward decomposition process that sequentially solves PRPs to identify the precursor set $\mathcal{P}$ for a given target T. Starting from T, each material is split into a reactant pair $\mathbf{p} = (p_1, p_2)$, where each p_i either serves as a final precursor or acts as an intermediate that is recursively decomposed in a subsequent step. The process continues until all materials have been reduced to final precursors, yielding $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$. For example, given the target $LiFeV_2O_7$, the first PRP step yields the pair $(LiVO_3,\ Fe_2O_3)$. In the next step, $LiVO_3$ is further decomposed into $(Li_2CO_3,\ V_2O_5)$ while Fe_2O_3 is retained as a final precursor. The resulting precursor set is $\mathcal{P} = \{Li_2CO_3,\ V_2O_5,\ Fe_2O_3\}$.

4. Methods

Retro-Forge (Figure 1) comprises three components: (i) a curated pairwise reaction dataset augmented with LLM-generated synthetic data to address data scarcity (Section 4.1), (ii) a pairwise reactant prediction model based on LLMs (Section 4.2), and (iii) a multi-step tree search that composes PRPs into complete synthesis routes (Section 4.3). The following sections describe each component in detail.

4.1. Pairwise Reaction Dataset and LLM-Based Augmentation

Pairwise Reaction Dataset. We construct a pairwise reaction dataset from two inorganic solid-state synthesis corpora (Kononova et al., 2019; Lee et al., 2025). Each source is preprocessed to resolve compositional variables and filtered for chemical validity (e.g., non-physical subscripts or coefficients). The two sources are merged under a common schema. Ambient species (O_2, N_2, CO_2, H_2O) are separated from solid reactants, and only reactions with exactly two solid reactants are retained. After deduplication by canonical element composition, 5,973 pairwise reactions across 4,476 unique targets remain, forming the Pairwise Reaction Dataset (PRD). Additional preprocessing details are provided in Appendix A.1.

LLM-Based Augmentation. To address the data scarcity of the PRD, we expand training coverage through a three-tier augmentation pipeline, where the PRD itself forms Tier 1 — the most rigorously verified and chemically valid pairwise reactions. A GPT-4.1-mini (OpenAI, 2025) model fine-tuned on Tier 1 data generates candidate pairwise reactions for two additional target pools of increasing scale but decreasing verification quality. For Tier 2, we use targets drawn from the original Kononova and Lee corpora (Kononova et al., 2019; Lee et al., 2025) that are absent from the PRD — these are materials with documented solid-state synthesis records but without verified pairwise reaction decompositions. For Tier 3, following the pipeline of Prein et al. (Prein et al., 2025b), targets are sampled from the Materials Project (Jain et al., 2013) using maximum-entropy sampling over embeddings encoded by MTEncoder (Prein et al., 2023), a multi-task pretrained transformer encoder for structure agnostic materials composition representation. This maximizes compositional diversity across chemical space but yields reactions that are purely model-generated and therefore noisier. Generated pairs are filtered for chemical validity, element consistency between the pair and the target, and absence from the PRD. This yields approximately 16,768 and 62,532 additional pairwise reactions for

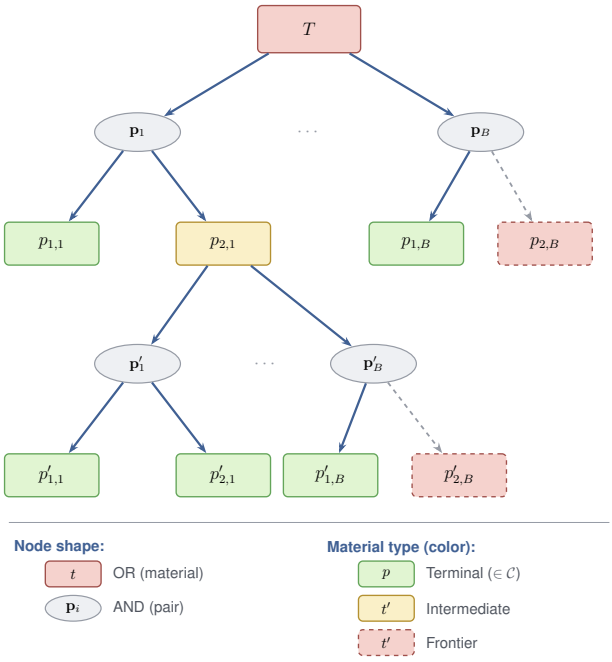

Figure 2. AND-OR tree structure of pairwise retrosynthesis. Rectangles are OR nodes (materials admitting B candidate pairs); ellipses are AND nodes (pairs whose two children must both be resolved). Color indicates material status: green = terminal ($\in \mathcal{C}$), amber = intermediate (recursively expanded), coral dashed = frontier (awaiting expansion).

13,336 targets of Tier 2 and 47,546 targets of Tier 3, respectively, giving a total training corpus of approximately 84,000 pairwise reactions across all three tiers.

4.2. Pairwise Reactant Prediction (PRP) Task

For the PRP task, we adopt large language models (LLMs) as the backbone for single-step pairwise reactant prediction. We evaluate four model variants: two locally-deployed open-source models, Qwen2.5-7B-Instruct (Qwen Team, 2025) and LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), and two API-based models, GPT-4o and GPT-4o-mini (OpenAI, 2024). All four models are fine-tuned on the PRD and its augmented extensions using a staged fine-tuning strategy that proceeds from large-scale noisy data to high-quality verified data: Tier 3 $\rightarrow$ Tier 2 $\rightarrow$ Tier 1 (PRD). This ordering exposes the model first to broad chemical space coverage before specializing on verified pairwise reactions. The locally-deployed models are fine-tuned using QLoRA (Dettmers et al., 2023), with two merging strategies explored per stage: continual tuning of the same LoRA adapter, or merging the previous stage's LoRA weights into the base model before the next stage. The API-based models are subjected to full fine-tuning. Training details and hyperparameters are provided in Appendix B.2.

Algorithm 1 Best-First Search over AND-OR Tree

Input: Target T, catalog $\mathcal{C}$, branching factor B, max expansions E, max depth D
Output: Top-k complete routes
Initialize $r_0 \leftarrow T$; $S(r_0) \leftarrow 0$; $\mathcal{Q} \leftarrow \{r_0\}$; $\mathcal{R} \leftarrow \emptyset$; $n \leftarrow 0$
while $\mathcal{Q} \neq \emptyset$ **and** $n < E$ **do**
  $r \leftarrow \mathcal{Q}$.PopMax()
  $t \leftarrow$ frontier node of r s.t. $t \notin \mathcal{C}$ and depth$(t) < D$
  if no such t exists **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{r\}$; **continue**
  end if
  $\{\mathbf{p}_i, s_i\}_{i=1}^{B} \leftarrow \text{PRP}(t, B)$
  for $i = 1$ **to** B **do**
    if $p_{1,i}, p_{2,i} \notin \text{Anc}(t, r)$ **then**
      Expand t into $p_{1,i}, p_{2,i}$ in $r \rightarrow r'$
      $S(r') \leftarrow S(r) + s_i$
      $\mathcal{Q}$.Push$(r', S(r'))$
    end if
  end for
  $n \leftarrow n + 1$
end while
return top-k from $\mathcal{R}$ by $S(\cdot)$

4.3. Precursor Prediction (PP) Task

For the PP task, we deploy the PRP models within a multi-step tree search that recursively applies pairwise reactant prediction until all leaf nodes correspond to purchasable precursors in the precursor catalog.

Precursor Catalog. We define the *precursor catalog* as the set of materials considered purchasable terminal nodes in the tree search. It is constructed from compounds appearing exclusively as precursors in the Kononova and Lee corpora (Kononova et al., 2019; Lee et al., 2025), restricted to those containing at most one metal or metalloid element excluding boron and silicon, and supplemented by entries from commercial chemical suppliers. A leaf node is considered terminal if its formula appears in the catalog, at which point no further decomposition is performed. Full catalog details are provided in Appendix A.3.

Tree Search. We formulate multi-step retrosynthesis as a best-first search over partial synthesis routes (Algorithm 1), following the AND-OR search framework used in organic retrosynthesis (Segler et al., 2018; Chen et al., 2020). Each route r is an AND-OR tree: a material that admits multiple candidate decompositions constitutes an OR node, while each specific decomposition $\mathbf{p} = (p_1, p_2)$ constitutes an AND node whose two children must both be resolved. Leaf materials in a route are classified as either *terminal* (present in the catalog $\mathcal{C}$) or *frontier* (not present in $\mathcal{C}$). A route is *complete* when all its leaves are terminal, and *partial* when at least one frontier material remains.

The search maintains a priority queue $\mathcal{Q}$ of partial routes ranked by cumulative log-probability $S(r)$. At each step, the highest-scoring partial route r is popped and a frontier material t is selected. The PRP model generates B candidates, PRP(t, B), each returning a reactant pair $\mathbf{p}_i = (p_{1,i}, p_{2,i})$ with sequence log-probability s_i. Each valid candidate—one in which neither $p_{1,i}$ nor $p_{2,i}$ appears among the ancestors of t in r (cycle check)—produces a new partial route r' by replacing t with its two children, with updated score $S(r') = S(r) + s_i$. Each child is marked terminal if it belongs to $\mathcal{C}$, or frontier otherwise. The search terminates when a complete route is found (added to $\mathcal{R}$), or when the expansion budget E is exhausted. The top-k complete routes by $S(\cdot)$ are returned. Search hyperparameters are given in Appendix B.3. The AND-OR tree structure is illustrated in Figure 2.

4.4. Evaluation Setup

Target-wise Top-k metric. Given a target T with ground-truth precursor sets $\{\mathcal{P}_1^*, \mathcal{P}_2^*, \ldots\}$, a prediction is considered correct at rank k if any one of the ground-truth sets is matched by any of the top-k predicted precursor sets $\hat{\mathcal{P}}_1, \ldots, \hat{\mathcal{P}}_k$. The metric is reported as the fraction of targets correctly predicted across the evaluation set at $k \in \{1, 3, 5, 10\}$, with matching performed at the stoichiometric level. We report results per target, in contrast to the reaction-wise top-k exact match used in prior work (Noh et al., 2025; Prein et al., 2025a), which evaluates each documented reaction record as a separate test instance. Under the reaction-wise protocol, a model given target T must exactly reproduce the specific precursor set recorded for that reaction, even when the same target appears multiple times in the corpus with different documented routes. This conflates the one-to-many nature of synthesis planning with a recall task over individual reaction records. Our target-wise evaluation instead asks once per target whether the model can identify any one valid precursor set, which is both chemically appropriate and consistent with the actual inference setting where the model receives only T as input.

Multi-step Evaluation Targets. To evaluate models on reactions outside the training distribution and beyond the pretraining cutoff of the language models used, we construct an evaluation set from papers registered in the ICSD with a 2025 entry date, corresponding to experimental works published after October 2024—beyond the pretraining cutoff of all language models used in this work. Relevant synthesis text from main and supplementary sections is manually identified for each paper and provided to Claude Sonnet 4.6 (Anthropic, 2026) for structured extraction. This process yields 110 PP evaluation targets after chemical validity filtering, including eight targets with multiple valid synthesis routes.

Table 1. Pairwise reactant prediction on Tier 1 test (target-wise top-k, %). All PRP models of Retro-Forge trained through full stages of Tier 3→2→1. **Bold** indicates the best performance, while underline represents the second best performance.

Model	@1	@3	@5	@10
MSP-LLM	67.6	77.5	81.1	83.3
LLaMA Continual	<u>80.4</u>	88.5	91.2	<u>94.6</u>
LLaMA Merge	**80.6**	**92.1**	**93.5**	**95.5**
Qwen Continual	79.9	86.9	91.2	93.9
Qwen Merge	80.2	<u>89.0</u>	<u>91.4</u>	<u>94.6</u>
GPT-4o	78.6	85.4	87.2	88.5
GPT-4o-mini	77.0	85.8	87.2	89.2

5. Experiments

We evaluate Retro-Forge based on multiple LLM families and training strategies for both the PRP and PP tasks, comparing against the state-of-the-art model on the PP task, MSP-LLM (Noh et al., 2026). Details of baseline model reproduction are shown in Appendix C. Dataset splits, training configurations, and inference details are provided in Appendix B.2.

5.1. Pairwise Reactant Prediction (PRP) Task

Setup. PRP models are evaluated on the Tier 1 test split (Appendix B.2), which contains 599 reactions over 444 unique targets held out from all augmented training data. For each target, the model generates up to 10 candidate reactant pairs ranked by sequence log-probability. We apply the target-wise top-k metric defined in Section 4.4: a target is correct at rank k if any of its ground-truth pairwise reactions is matched by any of the top-k predicted pairs, with matching at the stoichiometric level. We additionally include MSP-LLM as a reference, evaluated on the same test set with a prediction counted correct only if the predicted set exactly matches the two-precursor ground-truth pair; as MSP-LLM was not trained for the pairwise formulation, this serves as a reference point rather than a direct comparison.

Results. Table 1 reports PRP accuracy for four model families, each trained through the full Tier 3→2→1 staged fine-tuning. Both locally-deployed (Qwen, LLaMA) and API-based (GPT-4o, GPT-4o-mini) models achieve target-wise top-1 accuracy between 77% and 80%, indicating that the pairwise formulation is learnable across a range of model scales and families. MSP-LLM reaches 67.6%, notably lower, as expected given that it was not trained for the pairwise formulation.

5.2. Precursor Prediction (PP) Task

Each PRP model is deployed within the multi-step tree search to perform pairwise retrosynthesis. Evaluation is on the 110 PP evaluation targets using the target-wise top-k

Table 2. Precursor prediction on 110 post-Oct-2024 targets (target-wise top-k, %). **Bold** indicates the best performance, while underline represents the second best performance.

Model	@1	@3	@5	@10
MSP-LLM (Noh et al., 2026)	**40.9**	**50.9**	**53.6**	**56.4**
Qwen Continual	**40.9**	41.8	41.8	41.8
Qwen Merge	34.5	40.9	40.9	41.8
LLaMA Continual	30.0	37.3	43.6	48.2
LLaMA Merge	30.0	38.2	44.5	47.3
GPT-4o	39.1	40.9	42.7	44.5
GPT-4o-mini	39.1	42.7	45.5	46.4

Table 3. Route diversity in the multi-step search. Uniq@k = avg. distinct procurement lists in top-k routes. Share #1 = fraction of top-10 routes sharing the highest-ranked route's initial decomposition (not applicable to MSP-LLM, which predicts precursor sets in a single step without tree search).

Model	Uniq@3	Uniq@5	Uniq@10	Share #1
MSP-LLM	2.8	4.5	8.7	—
Qwen Cont.	1.8	2.6	4.8	37%
Qwen Merge	1.8	2.6	5.0	35%
LLaMA Cont.	1.8	2.7	4.4	39%
LLaMA Merge	2.1	3.0	5.0	33%
GPT-4o	1.8	2.4	3.4	45%
GPT-4o-mini	1.7	2.2	3.3	47%

metric.

Table 2 reports precursor prediction performance. For locally-deployed models, we compare the Continual and Merge QLoRA training strategies described in Appendix B.2. Retro-Forge with Qwen Continual (T3→2→1) matches MSP-LLM at top-1 (40.9%). GPT-4o and GPT-4o-mini both reach 39.1%, closely following. MSP-LLM retains an advantage at higher k, achieving best performance from @3 to @10. To understand this gap, we analyze the route diversity produced by the multi-step search (Table 3). Although each single-step expansion generates B=5 candidate pairs, the best-first priority queue preferentially expands the highest-scoring subtree: on average, 35–47% of a target's top-10 routes share the same initial pairwise decomposition as the top-ranked route. This structural bias toward depth over breadth limits the number of genuinely distinct precursor sets in the top-k (e.g., only 2.6 unique procurement lists at @5 for Qwen Continual despite 5 beams per expansion). Notably, models with higher top-1 accuracy show *less* top-k improvement (Qwen Continual: +0.9 pp from @1→@10), while models with lower top-1 show large gains (LLaMA Continual: +18.2 pp)—indicating that the limited improvement reflects ranking quality rather than route diversity.

Case study. Retro-Forge and MSP-LLM achieve comparable top-1 (40.9%) but succeed on *different* subsets of targets: of the 110 PP evaluation targets, 33 are correct under both, while 12 are solved only by Retro-Forge and the other 12 only by MSP-LLM.

The 12 Retro-Forge-only successes exploit multi-step decomposition through chemically valid intermediates. For example, the five-precursor target $Ba_2(MgCo)TeB_2O_{10}$ is recovered via a four-step route through the double perovskite Ba_2MgTeO_6, from $Co_2B_2O_5$ and $MgTeO_3$, as illustrated in Appendix D.3 (Shanmugapriya et al., 2025). Similarly, $Ca_2TeV_2O_9$ is decomposed through $Ca_2V_2O_7$ into ground truth precursors of $CaCO_3$, V_2O_5, and TeO_2 (Wang et al., 2025). These cases illustrate the core advantage of pairwise retrosynthesis: the ability to reason through intermediate compounds that connect purchasable precursors to the target via a sequence of chemically plausible steps.

Conversely, MSP-LLM's 12 unique successes predominantly involve elemental precursor routes and halide chemistry. For example, MSP-LLM correctly predicts the elemental set $\{Cu,\ S,\ Sr,\ Zr\}$ for $Sr_3Zr_2Cu_4S_9$ (Barman et al., 2025), whereas Retro-Forge defaults to binary compounds (Cu_2S, SrS, ZrS_2). Similarly, for the lithium holmium chlorobromide series $Li_3HoCl_{6-x}Br_x$ (x = 0–3), MSP-LLM identifies the correct halide precursors, $HoCl_3$, LiBr, and LiCl (Ogbolu et al., 2025), while Retro-Forge routes through the oxide Ho_2O_3 with NH_4Cl as a chlorine source—a plausible but incorrect synthesis strategy. These patterns suggest that MSP-LLM's explicit material group classification helps condition precursor selection on the target's chemical class, a mechanism that could be straightforwardly incorporated into the PRP model of Retro-Forge.

5.3. Data Augmentation Ablation

Table 4 ablates the contribution of each augmentation tier on both the PRP and PP tasks.

Pairwise Reactant Prediction Task. For locally-deployed models (Qwen, LLaMA), PRP top-1 increases monotonically with augmentation: Qwen improves from 72.1% (T1 only) to 79.9% (T3→2→1), and LLaMA from 78.4% to 80.4%. The augmented tiers expose the model to a broader range of pairwise reactions, improving generalization even on the in-distribution Tier 1 test set. API-based models (GPT-4o, GPT-4o-mini) show a different pattern: GPT-4o peaks at T1 only (81.8%) and slightly decreases with augmentation, suggesting that the stronger pretrained base already covers the chemical knowledge that augmentation provides to smaller models.

Precursor Prediction Task. On the PP evaluation targets, the full stages of fine-tuning consistently yield the highest top-1 across all model families, improving over T1 only trained models by 3.6–13.6 pp. This confirms that LLM-generated pairwise data substantially improves multi-step

Table 4. Dataset augmentation ablation (top-1, %). Left: pairwise reactant prediction on Tier 1 test. Right: precursor prediction on 110 PP evaluation targets. **Bold** indicates the best performance.

	PRP task (T1 Test)			PP task		
Model	**T1**	**T2→1**	**T3→2→1**	**T1**	**T2→1**	**T3→2→1**
Qwen Cont.	72.1	78.4	79.9	27.3	32.7	**40.9**
LLaMA Cont.	78.4	80.2	80.4	25.5	29.1	30.0
GPT-4o	**81.8**	77.7	78.6	33.6	33.6	39.1
GPT-4o-mini	75.7	76.8	77.0	35.5	34.5	39.1

precursor prediction, extending prior finding (Prein et al., 2025b) on single-step prediction to the multi-step setting.

6. Conclusion

Summary. We presented Retro-Forge, a multi-step pairwise retrosynthesis framework for solid-state materials synthesis planning. To the best of our knowledge, this is the first work to formulate precursor prediction for solid-state synthesis as a sequence of learnable pairwise reactions— directly addressing the fundamental limitation of prior approaches that treat precursor prediction as a single-shot mapping while overlooking the pairwise nature of solid-state reactions—and to evaluate model performance under a target-wise, contamination-free protocol.

We constructed the Pairwise Reaction Dataset (PRD), the first chemically curated dataset of solid-state pairwise reactions, filtered from two large synthesis corpora to enforce chemical validity. To address the inherent data scarcity of verified pairwise reactions, we developed a data augmentation pipeline comprising three tiers of datasets, in which an LLM fine-tuned on the PRD generates synthetic pairwise reactions for diverse target pools drawn from existing synthesis corpora and the Materials Project. This augmentation consistently improves both single-step pairwise reactant prediction accuracy and multi-step precursor prediction across all model families, extending prior findings on single-step prediction to the multi-step setting.

The single-step pairwise formulation proves to be learnable on a range of model scales and families, with locally-deployed and API-based LLM models achieving top-1 accuracy between 77% and 80% on the Tier 1 test set. When deployed within a multi-step tree search, the best Retro-Forge configuration (Qwen Continual, T3→2→1) matches the state-of-the-art MSP-LLM baseline at top-1 (40.9%) on our contamination-free post-October-2024 evaluation set. Crucially, Retro-Forge and MSP-LLM succeed on complementary subsets of evaluation targets: Retro-Forge uniquely recovers targets that require multi-step decomposition through chemically valid intermediates, while MSP-LLM has an advantage on targets involving elemental precursor routes and halide chemistry. This complementarity suggests that the two approaches are genuinely different in their reasoning strategies, and that ensemble or hybrid approaches may be a productive direction.

Limitation. The current system has two primary limitations. First, while the best Retro-Forge configuration matches MSP-LLM at top-1, MSP-LLM retains a substantial advantage from top-3 onward. Analysis of route diversity in Table 3 reveals that this gap stems from a structural bias in the best-first tree search toward depth over breadth, causing top-k routes to share the same initial decomposition and yielding few genuinely distinct precursor sets at higher k. Second, the pairwise approach faces challenges on targets with unusual precursor chemistry—elemental routes, halides, and complex multi-metal systems—where the PRD and its augmented extensions have limited coverage, constraining top-1 performance on these chemical families.

Future Work. These limitations point toward clear directions for future work. Multi-step route discovery can be improved through two complementary directions: diversifying the tree search strategy—for example through stochastic sampling or explicit breadth-encouraging mechanisms—and improving single-step pairwise reactant prediction accuracy, both of which directly translate to broader and higher-quality coverage of the precursor search space. Expanding PRD coverage to underrepresented chemical families—particularly halides, nitrides, and chalcogenides—and improving the quality of LLM-generated augmentation data are the most impactful levers for improving generalization across target types.

Retro-Forge demonstrates that domain knowledge about how solid-state reactions proceed—pairwise, sequentially, through intermediates—can be effectively encoded as a learnable unit and composed into complete synthesis routes. Incorporating synthesis conditions—temperatures, atmospheres, and annealing schedules—into the pairwise formulation would bring the framework closer to a complete materials synthesis planning system, an important step as AI-driven materials discovery continues to propose candidate materials at unprecedented scale and the need for experimentally actionable synthesis plans becomes ever more critical.

References

Anthropic. Claude Sonnet 4.6 (claude-sonnet-4-6). https://www.anthropic.com, 2026. Large language model used for structured data extraction.

Aykol, M., Montoya, J. H., and Hummelshøj, J. Rational Solid-State Synthesis Routes for Inorganic Materials. *Journal of the American Chemical Society*, 143(24):9244–9259, 2021. doi: 10.1021/jacs.1c04888.

Barman, S., Yadav, S., Ray, A. K., Swati, Deepa, M., Niranjan, M. K., and Prakash, J. $Sr_3Zr_2Cu_4Q_9$ (Q = S and Se): two novel layered quaternary mixed transition metal chalcogenides. *Dalton Trans.*, 54:1871–1883, 2025. doi: 10.1039/D4DT02928C.

Chen, B., Li, C., Dai, H., and Song, L. Retro*: Learning retrosynthetic planning with neural guided A* search. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1608–1616, 2020. doi: 10.48550/arXiv.2006.15820.

Cheng, M., Luo, W., Tang, H., Yu, B., Cheng, Y., Xie, W., Li, J., Kulik, H. J., and Li, M. Enhancing Materials Discovery with Valence Constrained Design in Generative Modeling. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2507.19799.

Chung, V., Walsh, A., and Payne, D. J. Solid-state synthesizability predictions using positive-unlabeled learning from human-curated literature data. *Digital Discovery*, 4:2439–2453, 2025. doi: 10.1039/d5dd00065c.

Corey, E. J. and Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science*, 166:178–192, 1969. doi: 10.1126/science.166.3902.178.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. doi: 10.48550/arXiv.2305.14314.

Engel, E. and Dreizler, R. *Density Functional Theory: An Advanced Course*. Theoretical and Mathematical Physics. Springer Berlin Heidelberg, 2011. ISBN 9783642140907.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The Llama 3 Herd of Models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2407.21783.

He, T., Huo, H., Bartel, C. J., Wang, Z., Cruse, K., and Ceder, G. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science Advances*, 9(23):eadg8180, 2023. doi: 10.1126/sciadv.adg8180.

Huo, H., Bartel, C. J., He, T., Trewartha, A., Dunn, A., Ouyang, B., Jain, A., and Ceder, G. Machine-Learning Rationalization and Prediction of Solid-State Synthesis Conditions. *Chemistry of Materials*, 34(16):7323–7336, 2022. doi: 10.1021/acs.chemmater.2c01293.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

Kim, S., Noh, J., Gu, G. H., Chen, S., and Jung, Y. Predicting synthesis recipes of inorganic crystal materials using elementwise template formulation. *Chem. Sci.*, 15:1039–1045, 2024. doi: 10.1039/D3SC03538G.

Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., and Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6:203, 2019. doi: 10.1038/s41597-019-0224-1.

Lee, S., Cruse, K., Baibakova, V., Ceder, G., and Jain, A. Text-mined dataset of solid-state syntheses with impurity phases using Large Language Model. *Scientific Data*, 12:1969, 2025. doi: 10.1038/s41597-025-06222-y.

Magar, I. and Schwartz, R. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022. doi: 10.18653/v1/2022.acl-short.18.

Malkowski, T. F., Sacci, R. L., McAuliffe, R. D., Acharya, S. R., Cooper, V. R., Dudney, N. J., and Veith, G. M. Role of Pairwise Reactions on the Synthesis of $Li_{0.3}La_{0.57}TiO_3$ and the Resulting Structure–Property Correlations. *Inorganic Chemistry*, 60(19):14831–14843, 2021. doi: 10.1021/acs.inorgchem.1c01832.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Miura, A., Bartel, C. J., Goto, Y., Mizuguchi, Y., Moriyoshi, C., Kuroiwa, Y., Wang, Y., Yaguchi, T., Shirai, M., Nagao, M., Rosero-Navarro, N. C., Tadanaga, K., Ceder, G., and Sun, W. Observing and Modeling the Sequential Pairwise Reactions that Drive Solid-State Ceramic Synthesis. *Advanced Materials*, 33(24):2100312, 2021. doi: 10.1002/adma.202100312.

Noh, H., Lee, N., Na, G. S., and Park, C. Retrieval-Retro: Retrieval-based Inorganic Retrosynthesis with Expert Knowledge. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2410.21341.

Noh, H., Na, G. S., Lee, N., and Park, C. MSP-LLM: A Unified Large Language Model Framework for Complete Material Synthesis Planning. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2602.07543.

Ogbolu, B. O., Poudel, T. P., Dikella, T. N. D. D., Truong, E., Chen, Y., Hou, D., Li, T., Liu, Y., Gabriel, E., Xiong, H., Huang, C., and Hu, Y.-Y. Tailoring Ion transport in $Li_{3-3y}Ho_{1+y}Cl_{6-x}Br_x$ via Transition-Metal Free Structural Planes and Charge [c]arrier Distribution. *Advanced Science*, 12(7):2409668, 2025. doi: https://doi.org/10.1002/advs.202409668.

Okabe, R., West, Z., Chotrattanapituk, A., Cheng, M., Carrizales, D. C., Xie, W., Cava, R. J., and Li, M. Large language model-guided prediction toward quantum materials synthesis. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2410.20976.

OpenAI. GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024.

OpenAI. GPT-4.1. https://openai.com/index/gpt-4-1/, 2025. Accessed: 2025.

Prein, T., Pan, E., Doerr, T., Olivetti, E., and Rupp, J. L. MTENCODER: A Multi-task Pretrained Transformer Encoder for Materials Representation Learning. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=wug7i3O7y1.

Prein, T., Pan, E., Haddouti, S., Lorenz, M., Jehkul, J., Wilk, T., Moran, C., Fotiadis, M. P., Toshev, A. P., Olivetti, E., and Rupp, J. L. M. Retro-Rank-In: A Ranking-Based Approach for Inorganic Materials Synthesis Planning. *arXiv preprint*, 2025a. doi: 10.48550/arXiv.2502.04289.

Prein, T., Pan, E., Jehkul, J., Weinmann, S., Olivetti, E. A., and Rupp, J. L. M. Language Models Enable Data-Augmented Synthesis Planning for Inorganic Materials. *ACS Applied Materials & Interfaces*, 17(51), 2025b. doi: 10.1021/acsami.5c09621.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2412.15115.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023 of the Association for Computational Linguistics*, pp. 10776–10787, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5:1572–1583, 2019. doi: 10.1021/acscentsci.9b00576.

Segler, M. H. S., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555:604–610, 2018. doi: 10.1038/nature25978.

Shanmugapriya, I. G., Sa, S., and Natarajan, S. Synthesis, structure, oxygen evolution reaction (OER) and visible-light assisted organic reaction studies on $A_2M_2TeB_2O_{10}$ (A = Ba and Pb; M = Mg, Zn, Co, Ni, Cu, and Fe). *Dalton Transactions*, 54:2753–2764, 2025. doi: 10.1039/D4DT02706J.

Song, Z., Lu, S., Ju, M., Zhou, Q., and Wang, J. Accurate prediction of synthesizability and precursors of 3D crystal structures via large language models. *Nature Communications*, 16(6530), 2025. doi: 10.1038/s41467-025-60875-0.

Sun, W. and David, N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discuss.*, 256:614–638, 2025. doi: 10.1039/D4FD00112E.

Szymanski, N. J. and Bartel, C. J. Computationally Guided Synthesis of Battery Materials. *ACS Energy Letters*, 9(6): 2902–2911, 2024. doi: 10.1021/acsenergylett.4c00821.

Szymanski, N. J., Nevatia, P., Bartel, C. J., Zeng, Y., and Ceder, G. Autonomous and dynamic precursor selection for solid-state materials synthesis. *Nature Communications*, 14(6956), 2023a. doi: 10.1038/s41467-023-42329-9.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., and Ceder, G. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624:86–91, 2023b. doi: 10.1038/s41586-023-06734-w.

Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., and Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Comput. Mater.*, 7 (1), 2021. doi: 10.1038/s41524-021-00545-1.

Wang, T., Luo, H., Zhang, S., Huang, L., Cao, L., Dong, X., and Zou, G. Enhanced Birefringence and Excellent Thermal Stability in Two Tellurium(IV) Vanadates: $KTeOVO_4$ and $Ca_2TeV_2O_9$. *Inorganic Chemistry*, 64:4673–4679, 2025. doi: 10.1021/acs.inorgchem.5c00306.

Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.*, 2(1), 2016. doi: 10.1038/npjcompumats.2016.28.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., Sordillo, R., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Yang, C., Li, W., Tomioka, R., and Xie, T. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5.

A. Dataset

A.1. Pairwise Reaction Dataset Construction

Our pairwise reaction dataset is constructed from two inorganic synthesis datasets (Kononova et al., 2019; Lee et al., 2025), containing 31,782 and 80,806 reactions, respectively. The two sources are preprocessed independently to accommodate their distinct structures, then unified under a common schema before pair extraction.

Source-Specific Preprocessing. For the dataset from Kononova et al., we resolve element and amount variables using the "element_substitution", "amounts_vars", and "targets_string" fields provided in the original metadata. For the dataset from Lee et al., element substitutions are parsed from inline annotations in the reaction strings. Entries with variables that cannot be resolved are discarded. In addition, this dataset stores multiple reactions per source publication in a single entry, which we flatten into individual entries.

Notational and Chemical Validity Filters. A series of filters are applied to remove entries with chemically invalid or computationally intractable formulas. We discard reactions with non-physical subscripts (negative or zero) and negative stoichiometric coefficients, as well as self-reactions whose reactant and product sets are identical. We also remove entries with disallowed characters or empty parentheses left by failed variable substitution. A further filter removes entries containing unresolved or unresolvable symbols in chemical formulas; these fall into three categories: (i) mathematical variables (x, y, δ) remaining after substitution failures, (ii) generic element placeholders referring to classes of elements (e.g., "Ln" for lanthanides), and (iii) parsing artifacts where a variable has been concatenated with an element symbol (e.g., "Wx", "Ok"). Abbreviated chemical groups such as isopropoxide ("i-OPr") are expanded to their full molecular formulas. Ambient species (O_2, N_2, CO_2, H_2O) are separated and removed from the solid reactants and products, as they do not serve as primary solid-phase reactants in solid-state synthesis. For example, a reaction such as $CaCO_3 \cdot H_2O + Co_3O_4 + O_2 \rightarrow Ca_3Co_4O_9 + CO_2$ is reduced to the solid-phase pair $(CaCO_3,\ Co_3O_4) \rightarrow Ca_3Co_4O_9$. Reactions in which the target contains elements absent from the precursors — excluding C, H, O, and N, which may be incorporated under ambient conditions — are removed. Reactions containing compounds with fractional or decimal subscripts are removed to prevent parsing errors. Entries are further filtered by charge neutrality to remove invalid chemical compounds.

Merging and Pair Extraction. The two preprocessed datasets are converted to a common schema and merged into a single reaction pool. From this pool, we retain reactions with exactly two solid reactants, as required by our pairwise formulation. We then deduplicate by constructing a canonical reaction key from the element-composition representation of both sides, which collapses notational variants of the same chemistry (e.g., TiO_2 and Ti_1O_2) into a single entry. When multiple entries share the same key, a single representative is retained.

After all processing steps, the final pair dataset contains 5,973 pairs.

A.2. Charge Neutrality Filter

Charge neutrality is enforced as a chemical validity filter at multiple stages of the pipeline: during PRD construction, during augmented data filtering, and during precursor catalog construction. The filter labels each compound formula with oxidation states using a BFS algorithm adapted from CrysVCD (Cheng et al., 2025), and accepts the compound only if a charge-neutral assignment exists. Compounds that cannot be labeled are discarded as chemically invalid.

Oxidation State Vocabulary. The algorithm operates over a vocabulary of 269 (element, oxidation state) tokens, covering common oxidation states for main-group elements, transition metals, lanthanides (La–Lu), and select actinides (Th, U, Np, Pu, Am). Each token corresponds to a specific ionic species (e.g., Fe^{2+}, Fe^{3+}, O^{2-}). A compound's oxidation state assignment maps each atom to one of its element's allowed tokens.

Two-pass labeling. Given a formula with element counts $\{(e_i, c_i)\}$, the algorithm attempts to assign oxidation states in two passes:

1. **Pass 1 (simple):** Assigns one oxidation state per element. The algorithm performs BFS over elements, trying each allowed oxidation state v for element e_i and accumulating a running charge $Q = \sum_i c_i \cdot v_i$. Branches are pruned when the last element's oxidation state has the same sign as the running charge (and thus cannot neutralize it). The formula is accepted if any leaf achieves $|Q| = 0$.
2. **Pass 2 (mixed-valence):** If Pass 1 fails, the algorithm expands each individual atom rather than each element, allowing

different oxidation states for the same element within one compound. For example, Fe_3O_4 requires both Fe^{2+} and Fe^{3+}. Consistency constraints prevent assigning both positive and negative oxidation states to the same element. The search is capped at 10,000 candidate compositions to bound computational cost.

Alloys (compounds where every element has 0 as an allowed oxidation state) are labeled directly with neutral oxidation states and bypass both passes.

Neutrality criterion. A formula is accepted if and only if the absolute charge residual is zero: $|Q| = 0$ (within floating-point tolerance 10^{-8}). No relaxed threshold is applied. Compounds with nonzero charge residual under all possible oxidation state assignments are rejected as chemically invalid. This strict criterion eliminates compounds with impossible or ambiguous oxidation state chemistry.

A.3. Dataset Analysis

Pairwise Reaction Dataset (PRD). The PRD comprises 5,973 pairwise reactions over 4,476 unique target materials and 1,109 unique precursors drawn from the Kononova (Kononova et al., 2019) and Lee (Lee et al., 2025) corpora. Table 5 summarizes the key statistics across all tiers. The PRD covers 81 elements and is dominated by oxides (46.0%) and alloys (35.0%). Most targets (84.5%) appear in only one reaction, and the average number of elements per target is 2.7. Of the 1,109 unique precursors, 787 (71.0%) are present in the precursor catalog, while the remaining 322 (29.0%) serve as intermediates in multi-step chains.

A connectivity analysis reveals that 287 products (6.4%) also appear as precursors in other reactions, indicating the presence of implicit synthesis chains within the PRD. Computing synthesis depth from the catalog, 89.1% of PRD targets are reachable in a single step (both precursors directly purchasable), 8.3% require two steps, and only 0.2% require three or more. This shallow depth profile explains why models trained on the PRD alone achieve high PRP accuracy but struggle on out-of-distribution targets that require deeper decomposition.

The most frequently used precursors are common oxide and carbonate reagents: TiO_2 (214 reactions), $SrCO_3$ (194), Li_2CO_3 (175), and Fe_2O_3 (174)—all present in the catalog. Precursor pair reuse is moderate: 27.9% of unique pairs produce more than one target (e.g., Fe_2O_3 + $SrCO_3$ $\rightarrow$ 8 different strontium ferrites).

Augmented PRD. Tier 2 adds 19,195 filtered reactions over 13,336 targets from the Kononova and Lee corpora, and Tier 3 adds 70,199 filtered reactions over 47,546 Materials Project targets. The augmented tiers substantially expand compositional diversity: Tier 2 shifts toward higher-complexity targets (average 3.3 elements/target vs. 2.7 for the PRD) with increased representation of sulfides (4.3% vs. 2.7%), selenides (3.3% vs. 1.2%), and phosphates (4.6% vs. 3.0%). Tier 3 further diversifies the distribution, with alloys comprising 58.8% of targets (reflecting the Materials Project's broad coverage of intermetallic phases) and introducing substantial nitride (2.7%) and carbide (2.7%) content absent from the PRD.

Critically, the augmented data introduces far more intermediate compounds: only 10.8% of Tier 2 precursors and 3.0% of Tier 3 precursors appear in the catalog, compared to 71.0% for the PRD. This means the augmented tiers primarily teach the model about *intermediate* chemistry rather than terminal precursor selection. Tier 2 in particular exhibits the richest multi-step structure: 12.9% of its products also appear as precursors in other reactions (vs. 6.4% for the PRD), and 29.3% of its targets require depth-2 synthesis from the catalog. This explains the consistent benefit of including Tier 2 in the dataset (Section 5.3).

Element coverage saturates early: the PRD covers 81 elements, Tier 2 adds 1, and Tier 3 adds 0. The augmentation benefit is therefore not element coverage but *compositional diversity* —new combinations of known elements in underrepresented chemical families.

Precursor Catalog. The precursor catalog contains 1,186 purchasable starting materials, constructed from compounds that appear exclusively as precursors (never as targets) in the Kononova and Lee corpora. The catalog spans 82 elements and comprises 94 elemental entries (7.9%), 569 binary compounds (48.0%), 363 ternary compounds (30.6%), and 124 quaternary or higher (10.5%).

By material category, oxides dominate (42.1%), followed by alloys (15.9%), halides (14.9%), sulfides (6.9%), phosphates (6.3%), and selenides (5.3%). The catalog's halide coverage (14.9%) is notably higher than its share in any training tier, reflecting the prevalence of halide salts as common laboratory reagents (e.g., LiF, NaCl, KBr).

Precursor Prediction Evaluation Targets. The PP evaluation set comprises 110 target materials from papers published

Table 5. Dataset statistics across all tiers, the precursor catalog, and the PP evaluation targets. "In catalog" indicates the fraction of unique precursors present in the catalog (for tiers) or of ground-truth precursors (for PP evaluation targets). Category percentages are computed over unique targets (or entries for the catalog).

	Tier 1 (PRD)	Tier 2	Tier 3	Eval Targets	Catalog
Reactions / entries	5,973	19,195	70,199	110	1,186
Unique targets	4,476	13,336	47,546	110	—
Unique precursors	1,109	7,041	28,265	151	1,186
Elements covered	81	82	82	71	82
Avg. elem./target	2.7	3.3	3.3	3.7	—
In catalog (%)	71.0	10.8	3.0	89.4	100
Oxide (%)	46.0	42.8	16.5	20.9	42.1
Alloy (%)	35.0	35.5	58.8	30.0	15.9
Phosphate (%)	3.0	4.6	1.0	3.6	6.3
Sulfide (%)	2.7	4.3	5.5	12.7	6.9
Selenide (%)	1.2	3.3	4.7	13.6	5.3
Halide (%)	5.2	2.3	5.6	5.5	14.9
Oxyhalide (%)	1.8	2.2	2.5	10.0	1.0
Other (%)	5.0	5.1	5.4	3.6	7.6

Table 6. Connectivity and synthesis depth statistics. Depth is computed by tracing existing pairwise reactions back to the precursor catalog. Targets are "unreachable from catalog" when no chain of reactions within the dataset connects catalog precursors to the target.

	T1	T2	T3	All
Unique pairs	4,003	15,535	59,887	72,507
Pairs reused (>1 target)	27.9%	13.7%	10.7%	12.8%
Products as precursors	6.4%	12.9%	6.3%	9.0%
Depth 1 (direct)	89.1%	34.1%	9.8%	8.9%
Depth 2	8.3%	29.3%	4.6%	8.0%
Depth 3+	0.2%	2.4%	1.6%	3.5%
Unreachable from catalog	2.4%	34.3%	84.1%	79.6%

after October 2024, ensuring temporal separation from all model training cutoffs. These targets are compositionally more complex than the PRD (average 3.7 elements/target; 45.5% have 4 elements, 15.5% have 5 or more) and exhibit a markedly different category distribution: selenides (13.6%), oxyhalides (10.0%), and sulfides (12.7%) are substantially overrepresented compared to the oxide-dominated PRD.

Of the 151 unique ground-truth precursors across all evaluation targets, 135 (89.4%) are present in the catalog. The 16 missing precursors include specialized compounds such as $AgBF_4$, Li_2NCN, NbO_2F, TaO_2F, $HoCl_3$, and $Ba(BF_4)_2$—targets whose ground-truth routes require these precursors are inherently unsolvable by the multi-step search regardless of model quality.

B. Implementation Details

B.1. Data Augmentation Details

GPT-4.1-mini Fine-Tuning. The data augmentation model is a fine-tuned `gpt-4.1-mini-2025-04-14` trained on the Tier 1 (PRD) training split (5,375 pairwise reactions) via the OpenAI fine-tuning API. The model is trained for 3 epochs (batch size 10, learning rate multiplier 2.0, 1.60M trained tokens; auto-selected by the OpenAI API) using the same chat-format prompt template as all downstream single-step PRP models (system prompt + "Predict two solid-state precursors or intermediates for `[target]`" → "`P1 + P2`"). The fine-tuned model is used exclusively for data generation and is *not* used at inference time.

Tier 2 Target Selection. Tier 2 targets are drawn from the Kononova (Kononova et al., 2019) and Lee (Lee et al., 2025) corpora. We select all targets that appear in these corpora but are absent from the PRD—materials with documented solid-state synthesis records but without verified pairwise reaction decompositions. After removing targets that overlap with the Tier 1 test split to prevent data contamination, 13,909 unique targets remain.

Tier 3 Target Selection. Tier 3 targets are sampled from the Materials Project (Jain et al., 2013) via maximum-entropy sampling over MTEncoder embeddings (Prein et al., 2025b). Candidate compositions are first filtered for integer stoichiometry, maximum element count ≤ 20, and charge neutrality. All Tier 1 test targets are excluded. From the filtered pool, 50,000 targets are selected to maximize compositional diversity across chemical space, covering underrepresented regions not present in the Kononova and Lee corpora.

Inference and Filtering. For each Tier 2 and Tier 3 target, the fine-tuned GPT-4.1-mini generates three candidate pairwise reactions. Generated pairs are filtered through three criteria before inclusion in the training set:

1. **Chemical validity:** both predicted formulas must be parseable as valid compositions with integer stoichiometry.
2. **Element consistency:** the union of elements in the predicted pair must cover all elements in the target, allowing for common byproduct-forming species (C, H, O, N from carbonates, hydroxides, and nitrates).
3. **Deduplication:** the generated pair must not already appear in the PRD (Tier 1) under canonical element-composition matching.

Table 7 summarizes the generation and filtering statistics.

Table 7. Data augmentation statistics. "Generated" = raw LLM output; "Filtered" = after chemical validity, element consistency, and deduplication filters; "Train" = final training split size.

Tier	Targets	Pairs (Generated)	Targets (Filtered)	Pairs (Filtered)
Tier 1 (PRD)	4,476	—	—	5,973
Tier 2	13,909	41,730	13,336	19,195
Tier 3	50,000	149,998	47,546	70,199

After filtering, the Tier 2 and Tier 3 data are split into training and validation sets with the constraint that no target appearing in the Tier 1 test split may appear in any tier's training data. The final training corpus comprises approximately 84,000 pairwise reactions across all three tiers (Tier 1: 4,720; Tier 2: 16,768; Tier 3: 62,532).

B.2. Pairwise Reactant Prediction Models Training Details

Dataset Splits. The PRD is split at the target level (seed 42) into 80/10/10% train/val/test. Tier 2 and Tier 3 are split into train and val only; all targets overlapping with the Tier 1 test set are excluded to prevent contamination. Table 8 summarizes the split sizes.

Table 8. Dataset split sizes (reactions / unique targets).

	Train		Val		Test	
	Rxns	Tgts	Rxns	Tgts	Rxns	Tgts
Tier 1 (PRD)	4,720	3,586	654	447	599	444
Tier 2	16,768	11,610	1,870	1,289	—	—
Tier 3	62,532	42,334	6,980	4,703	—	—

Locally-Deployed Models Fine-tuning Details. Both Qwen2.5-7B-Instruct (Qwen Team, 2025) and LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) models are fine-tuned using Python 3.10, PyTorch, Accelerate, and the HuggingFace TRL/PEFT stack on NVIDIA H100 80GB GPUs. We adopt QLoRA (Dettmers et al., 2023) for parameter-efficient fine-tuning to improve training efficiency under limited resources. To evaluate the effectiveness of knowledge acquisition, we compare two different methods of implementing LoRA adapters per stage:

- **Continual QLoRA** — A single LoRA adapter is sequentially reused across all three tiers. After each tier's stage completes, the same adapter is re-loaded and training continues. There is no weight merging between stages, allowing the adapter to evolve continuously.
- **Merge QLoRA** — Each tier trains a fresh LoRA adapter on top of the current base model. Upon completion of a tier, the adapter's updates are merged into the base weights before proceeding to the next tier. The subsequent tier then starts from this updated base and trains a new adapter from scratch. This approach is designed to explicitly freeze and consolidate earlier-tier knowledge.

Table 9. Hyperparameter configuration for Continual and Merge QLoRA fine-tuning.

Method	LoRA	Tier 3			Tier 2			Tier 1		
	r/α	ep	lr	bs	ep	lr	bs	ep	lr	bs
Continual LoRA	16 / 32	1	$2e-4$	2	2	$1e-4$	2	3	$5e-5$	2
Merge LoRA	16 / 32	3	$1e-4$	2	3	$1e-4$	2	3	$1e-4$	2

API-Based Models Fine-Tuning Details.

We fine-tune GPT-4o (`gpt-4o-2024-08-06`) and GPT-4o-mini (`gpt-4o-mini-2024-07-18`) via the OpenAI fine-tuning API. Both cutoffs are strictly before all 110 PP evaluation targets (published after October 2024), ensuring the same out-of-distribution guarantee as our locally-deployed models.

Each model is trained with the same three-tier stages as Qwen and LLaMA (Tier 3→2→1), with each stage continuing from the previous stage's fine-tuned checkpoint. We also train Tier 1-only and Tier 2→1 ablation variants. The training data format (chat-format JSONL with the pairwise system prompt) is identical across all model families. Table 10 lists the hyperparameters for both models. Batch size and learning rate multiplier are auto-selected by the OpenAI API; the API assigns a slightly lower learning rate multiplier to GPT-4o-mini (1.8 vs. 2.0).

Table 10. GPT fine-tuning hyperparameters per stage. Batch size and learning rate multiplier are auto-selected by the OpenAI API.

Model	Stage	Epochs	Batch	LR mult.	Tokens
GPT-4o	Tier 3	1	41	2.0	6.26M
GPT-4o	Tier 2	2	22	2.0	3.40M
GPT-4o	Tier 1	3	9	2.0	1.41M
GPT-4o-mini	Tier 3	1	41	1.8	6.26M
GPT-4o-mini	Tier 2	2	22	1.8	3.40M
GPT-4o-mini	Tier 1	3	9	1.8	1.41M

B.3. Multi-Step Task Tree-Search Details

Beam Search in Locally-Deployed Models. For locally-deployed models (Qwen2.5-7B-Instruct and LLaMA-3.1-8B-Instruct), which are run on local hardware using the HuggingFace `transformers` library, beam search is applied natively at the single-step level to generate B candidate pairwise decompositions for each input composition. Given a target composition t, the model autoregressively decodes a sequence of tokens representing a precursor pair $\mathbf{p} = (p_1, p_2)$ using the prompt template described in Appendix E. Beam search maintains the top-B partial sequences at each decoding step, ranked by cumulative log-probability, and returns B complete hypotheses at the end of generation. Duplicate pairs under canonical element-composition matching are removed, retaining the highest-scoring instance.

The sequence score for a generated pair $\mathbf{p} = (p_1, p_2)$ given target t is the sum of per-token log-probabilities:

$$s(\mathbf{p} \mid t) = \sum_{i=1}^{L} \log P_\theta(\texttt{tok}_i \mid \texttt{tok}_{<i}, t) \tag{1}$$

where L is the total number of output tokens and P_θ is the fine-tuned model. This score is used both to rank the B candidates from a single expansion and to accumulate path scores during multi-step tree search (Section 4.3). We use $B = 5$ throughout all experiments, with no length penalty (`length_penalty=1.0`), consistent with the log-probability scoring used in the tree search priority queue.

Repeated Sampling for API-Based Model. Since the OpenAI API does not expose beam search, we use stochastic sampling (n=5, temperature 1.0) with `logprobs=True` for multi-step tree search inference. The length-normalized sum of per-token log-probabilities serves as the sequence score for priority-queue ranking. Duplicate sampled pairs are deduplicated, retaining the highest-scoring instance. Temperature 1.0 is chosen to maximize unique candidate diversity at n=5, approximately matching the number of distinct candidates produced by beam search with B=5. All other search parameters match the locally-deployed configuration.

Multi-Step Tree Search Integration. The single-step pairwise reactant prediction serves as the node expansion operator within the multi-step tree search. At each expansion step, the PRP model is called on the current non-terminal node — a composition t absent from the precursor catalog — to generate $B{=}5$ candidate pair reactants $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_B$. Each candidate $\mathbf{p}_i = (p_1, p_2)$ extends the current partial route, and the cumulative route score is updated as:

$$S(\text{route}) = \sum_{\mathbf{p} \in \text{route}} s(\mathbf{p} \mid t) \tag{2}$$

where the sum is over all pairwise decomposition steps along the route, each contributing the log-probability score of predicting the reactant pair $\mathbf{p}$ given the corresponding intermediate target t. The priority queue selects the highest-scoring partial route for expansion at each step. The search terminates when a complete route — all leaves present in the precursor catalog — is found, or the maximum expansion budget is reached. The top-k complete routes by cumulative score are returned and evaluated at $k \in \{1, 3, 5, 10\}$.

Table 11. Tree search hyperparameters used in all multi-step experiments.

Parameter	Value
Beam width B (locally-deployed models)	5
Samples n (API-based models)	5
Sampling temperature (API-based models)	1.0
Max tree expansions	200
Max route depth	5
Returned complete routes	up to 10
Cycle detection	enabled

C. MSP-LLM Baseline

We used MSP-LLM (Noh et al., 2026) based on LLaMA-3.1-8B (Grattafiori et al., 2024) as the state-of-the-art baseline for the PP task. The model is re-implemented under our identical training infrastructure to ensure a fair comparison: the same QLoRA configuration, batch size, and learning rate schedule are applied. To ensure data parity, we use the publicly released training and validation splits without any modification. Our re-implementation achieves 71.32% top-1 accuracy on the released test set, within 1% of the reported 70.75% (LLaMA-3.1-8B, Split 1), confirming faithful reproduction.

Table 12. Hyperparameter configuration for the MSP-LLM baseline

Method	LLM	Data	LoRA (r/α)	ep	lr	bs	patience
MSP-LLM	Llama-3.1-8B	Noh et al. (2026)	16 / 32	3	$1e-4$	2	3

D. Supplementary Results

D.1. Target-wise and Reaction-wise Top-k

The Tier 1 test set contains 599 reactions over 444 unique targets. Since multiple reactions can share the same target (with different ground-truth precursor pairs), the conventional *reaction-wise* (each of the 599 reactions scored independently) and our *target-wise* (a target is correct if any of its reactions matches at top-k) metrics differ. Here we report both values to give a comparison between them. Target-wise scores are consistently higher because a target with multiple valid precursor routes has more chances to match.

Table 13 shows both metrics for the Qwen Continual and Merge models across all top-k thresholds. The gap between reaction-wise and target-wise is approximately 20 pp at top-1 and narrows to $\sim$9 pp at top-10, reflecting that the top-k expansion recovers additional valid routes for multi-route targets.

D.2. Catastrophic Forgetting Analysis

Sequential staged training risks catastrophic forgetting: as the model specializes on later tiers, knowledge from earlier tiers may be overwritten. We assess forgetting by evaluating each model's pairwise prediction accuracy on the Tier 3 validation

Table 13. Reaction-wise vs. target-wise pairwise accuracy on Tier 1 test (stoichiometry matching, %). 599 reactions, 444 unique targets.

Model	Metric	@1	@3	@5	@10
Qwen Cont.	Reaction-wise	59.4	72.1	77.3	86.0
	Target-wise	80.2	88.7	91.7	95.0
Qwen Merge	Reaction-wise	59.4	72.5	77.3	85.8
	Target-wise	80.2	89.0	91.4	94.6

set at the end of each fine-tuning stage: after Tier 3 training (before any fine-tuning on Tier 2 or 1), after Tier 2, and after Tier 1.

Observing Forgetting. Table 14 tracks Tier 3 val accuracy as training progresses through the T3→T2→T1 stages for both locally-deployed and API-based models under the Continual and Merge training strategies.

Table 14. Tier 3 val pairwise accuracy (target-wise top-1, %) measured after each stage. A decreasing trend indicates forgetting; an increasing trend indicates continued learning.

Model	Strategy	After T3	After T2	After T1
Qwen	Continual	54.8	52.0	50.8
Qwen	Merge	55.3	53.4	52.1
LLaMA	Continual	35.0	42.2	56.9
LLaMA	Merge	49.2	46.8	56.3
GPT-4o	Sequential FT	52.9	52.8	51.1
GPT-4o-mini	Sequential FT	52.2	51.3	48.6

Qwen and GPT models exhibit mild but consistent forgetting: Tier 3 val accuracy drops 2–4 pp from after-T3 to after-T1, with Merge strategies retaining slightly more knowledge than Continual in both cases. In contrast, LLaMA shows the opposite pattern—Tier 3 val accuracy *increases* as training progresses (35.0% → 56.9% for Continual), indicating that LLaMA's base model does not fully acquire Tier 3 chemistry from T3 training alone and continues to improve through subsequent stages.

Mitigating Forgetting. To address the forgetting observed in Qwen and GPT models, we evaluate experience replay strategies on Qwen-2.5-7B, mixing a random subset of previous tiers' training data into each subsequent tier's training set. We test replay ratios of 5%, 10%, and 20%, as well as a variant combining 10% replay with a reduced Tier 1 learning rate (10^{-5} instead of 5×10^{-5}).

Table 15 reports the results. None of the replay variants improve over the Qwen Continual baseline (40.9% top-1). The best variant (Replay 10% + Low LR) matches the baseline exactly at top-1, while providing marginal gains at higher k (up to +1.8 pp at top-5). Other replay ratios slightly underperform at top-1 (40.0%). This suggests that Qwen's observed forgetting (4.0 pp drop on Tier 3 val) does not meaningfully harm downstream PP performance—the model retains sufficient earlier-tier knowledge despite the measured accuracy decline.

Table 15. Anti-forgetting strategies applied to Qwen: precursor prediction on 110 evaluation targets (target-wise top-k, %). No strategy improves over the Continual baseline.

Strategy	@1	@3	@5	@10
Qwen Continual (baseline)	**40.9**	41.8	41.8	41.8
Replay 10% + Low LR	**40.9**	**42.7**	**42.7**	**43.6**
Replay 10%	40.0	**42.7**	**44.5**	**44.5**
Replay 5%	40.0	**42.7**	42.7	43.6
Replay 20%	40.0	**42.7**	42.7	**45.5**

D.3. Case study: multi-step pairwise retrosynthesis of $Ba_2(MgCo)TeB_2O_{10}$

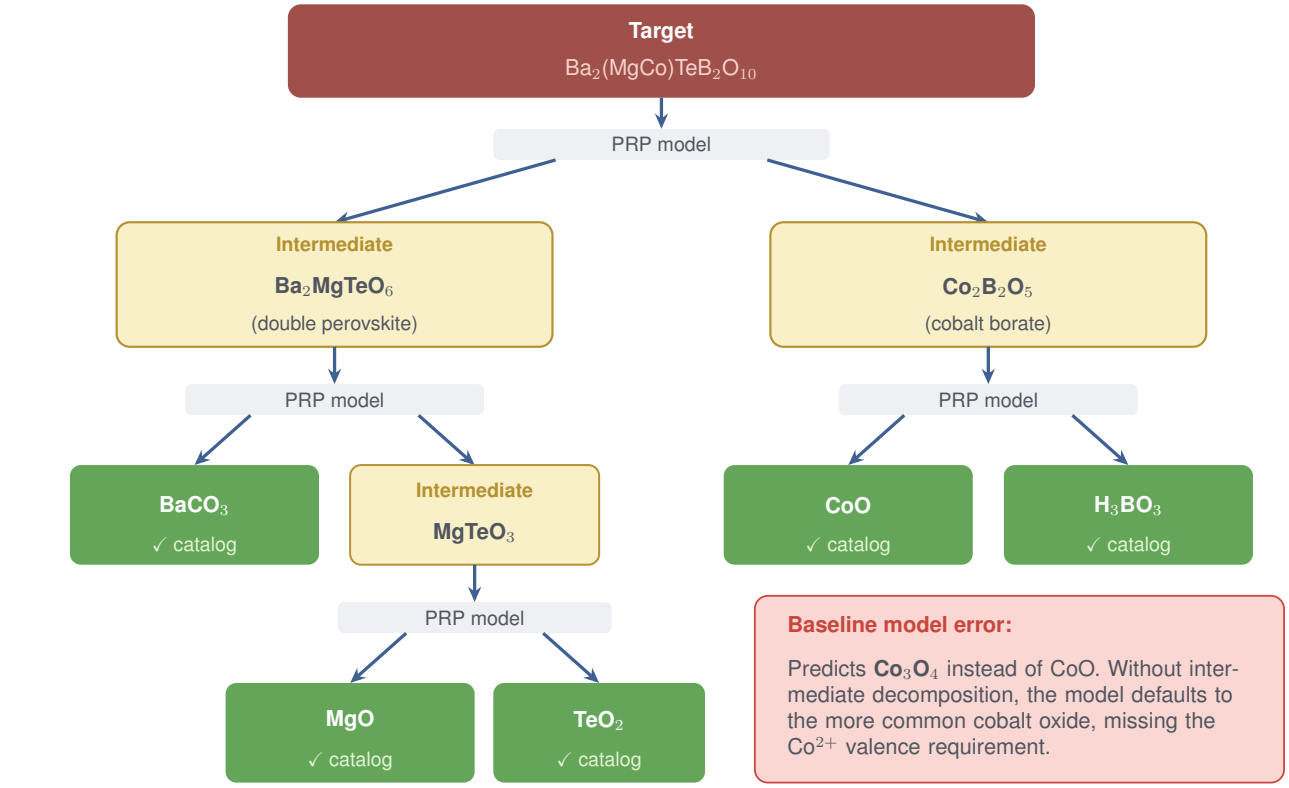

Figure 3. Case study: multi-step pairwise retrosynthesis of $Ba_2(MgCo)TeB_2O_{10}$, a 5-precursor target from the PP evaluation set. Retro-Forge decomposes the target through a sequence of pairwise reactions, producing intermediates Ba_2MgTeO_6 (double perovskite) and $Co_2B_2O_5$ (cobalt borate). At each decomposition step, the PRP model selects precursors consistent with the intermediate's local chemistry—correctly choosing CoO (Co^{2+}) over the more common Co_3O_4 (mixed Co^{2+}/Co^{3+}). All five predicted leaf precursors exactly match the ground truth (Shanmugapriya et al., 2025). A baseline model predicting all precursors in one shot defaults to Co_3O_4, lacking the intermediate-level charge balance constraint that guides correct selection.

E. Prompt Templates

In this section, we present the prompt templates used for the single-step model and data augmentation. All models share the same system prompt and user prompt; only the output format differs between model families.

Table 16. Prompt for Qwen, GPT-4o, GPT-4o-mini (PRP models) and GPT-4.1-mini (data augmentation model).

System Prompt: You are an expert in solid-state inorganic synthesis. Assume all solid-state inorganic synthesis happens pairwise between intermediates or precursors. Given a target material formula, predict exactly two precursor or intermediate materials that reacts pairwise via a solid-state reaction to synthesize the target. Output only the two material formulas separated by `' + '`.

User: Predict two solid-state precursors or intermediates for `{target_formula}`

Output format:

```
precursor1 + precursor2
```

Table 17. Prompt for LLaMA (PRP model).

System Prompt: You are an expert in solid-state inorganic synthesis. Assume all solid-state inorganic synthesis happens pairwise between intermediates or precursors. Given a target material formula, predict exactly two precursor or intermediate materials that reacts pairwise via a solid-state reaction to synthesize the target. Output only a JSON list of the two material formulas, e.g. `["A", "B"]`.

User: Predict two solid-state precursors or intermediates for `{target_formula}`

Output format:

```
["precursor1", "precursor2"]
```