RETRO-RANK-IN: A RANKING-BASED APPROACH FOR INORGANIC MATERIALS SYNTHESIS PLANNING

Thorben Prein[1,2,3,4], Elton Pan[5], Sami Haddouti[1,3], Marco Lorenz[1,3], Janik Jehkul[1,3], Tymoteusz Wilk[1,3], Cansu Moran[1,3], Menelaos Panagiotis Fotiadis[1,3], Artur P. Toshev[1], Elsa Olivetti[5], Jennifer L.M. Rupp[1,4]

[1]Technische Universität München, [2]Munich Data Science Institute, [3]TUM.ai, [4]TUMint. Energy Research GmbH, [5]Massachusetts Institute of Technology

Correspondence to: Jennifer L.M. Rupp <jrupp@tum.de>

ABSTRACT

Retrosynthesis strategically plans the synthesis of a chemical target compound from simpler, readily available precursor compounds. This process is critical for synthesizing novel inorganic materials, yet traditional methods in inorganic chemistry continue to rely on trial-and-error experimentation. While emerging machine-learning approaches struggle to generalize to entirely new reactions due to their reliance on known precursors, as they frame retrosynthesis as a multi-label classification task. To address these limitations, we propose *Retro-Rank-In*, a novel framework reformulating the **Retro**synthesis problem by embedding target and precursor materials into a shared latent space and learning a pairwise **Rank**er on a bipartite graph of **In**organic compounds. We evaluate Retro-Rank-In's generalizability on challenging retrosynthesis dataset splits designed to mitigate data duplicates and overlaps. For instance, for Cr_2AlB_2, it correctly predicts the verified precursor pair CrB + Al despite never seeing them in training, a capability absent in prior work. Extensive experiments show that Retro-Rank-In sets a new state-of-the-art, particularly in out-of-distribution generalization and candidate set ranking, offering a powerful tool for accelerating inorganic material synthesis.

1 INTRODUCTION

The discovery of inorganic materials underpins a wide array of modern technologies, such as renewable energy and electronics. Recent efforts involving large-scale computational exploration of the materials chemical space (Sriram et al., 2024; Merchant et al., 2023; Kim et al., 2021; Zhu et al., 2024) have led to the discovery of millions of potentially stable and synthesizable compounds (*what* to synthesize) (Merchant et al., 2023; Barroso-Luque et al., 2024; Saal et al., 2013; Zeni et al., 2023; Schmidt et al., 2024). However, the synthesis of these novel materials remains a critical bottleneck (*how* to synthesize) (Karpovich et al., 2023; Mahbub et al., 2020; Malik et al., 2021). Unlike inorganic materials, organic molecules exist as discrete, individual structures, which allow their synthesis to be broken down into multiple steps, each with smaller building blocks through a well-understood sequence of reactions – a process called *retrosynthesis*. In contrast, inorganic materials adopt a

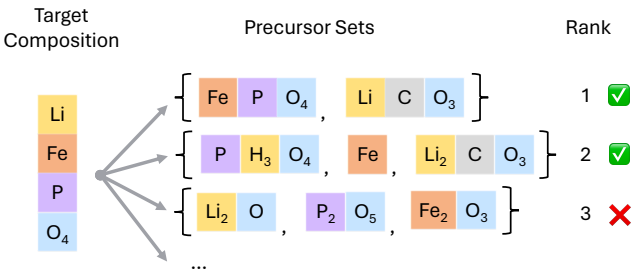

Figure 1: **Retrosynthesis problem.** Identifying the optimal precursor set for a given target material can be treated as a ranking problem. We use the binary classification probabilities of each set to determine its rank. Checkmarks indicate whether a ranked set corresponds to an experimentally verified synthesis.

periodic structure 3D arrangement of atoms. This periodicity renders the retrosynthetic strategy known in organic synthesis inapplicable to inorganic materials. The synthesis of inorganic materials largely remains a one-step process, where a *set* of precursors undergo a reaction to form a desired target compound. This complex process has no general, unifying theory (Kononova et al., 2019), and thus heavily relies on trial-and-error experimentation of precursor materials. Furthermore, the exponential scaling of compute needed for simulation impedes physical modeling of the underlying physical phenomena, e.g., thermodynamics and kinetics at the atomic scale (Bianchini et al., 2020).

This presents a compelling opportunity for machine learning (ML) approaches to bridge the knowledge gap by learning directly from synthesis data. In particular, precursor recommendation stands out as a key task in inorganic materials synthesis (Miura et al., 2021; Bianchini et al., 2020). For a reaction $A + B \rightarrow C$, the task is to recommend a set of precursors $\{A, B\}$ given target C. Early work in the field utilized a text-conditioned conditional variational autoencoder to generate synthesis precursors for novel materials (Kim et al., 2020). ElemwiseRetro (Kim et al., 2022) employs domain heuristics and a classifier for template completions. More recently, studies have leveraged language models to uncover and analyze relationships between target materials and their precursors, (Kim et al., 2024). An orthogonal approach trains a reaction template retriever by learning representations of target materials using a masked precursor completion task. These learned representations are then used to retrieve records of known syntheses of materials similar to the target material, which achieves strong performance in precursor recommendation (He et al., 2023).

Notably, the most recent work Retrieval-Retro (Noh et al., 2024) employs two retrievers, the first identifying reference materials sharing similar precursors with the target material, while the second retriever suggests precursors based on formation energies. Specifically, this approach uses self-attention and cross-attention for target-reference material comparison and predicts precursors via a multi-label classifier. This framework effectively unifies both a data-driven and a domain-informed approach for inorganic retrosynthesis. However, existing ML approaches face significant limitations, as summarized in Table 1. Most notably, they lack the ability to incorporate new precursors, a critical aspect of experimental workflows in laboratories when searching for novel precursors and discovering new compounds (McDermott et al., 2023; Szymanski & Bartel, 2024). For instance, Retrieval-Retro (Noh et al., 2024) cannot recommend precursors outside its training set, as they are represented through one-hot encoding in its multi-label classification output layer (Figure 2, a.). This design restricts the model to recombining existing precursors into new combinations rather than enabling predictions involving entirely novel precursors that have not been seen during training, thereby limiting its applicability in a material discovery setting. Furthermore, prior methods struggle to effectively incorporate broader chemical knowledge. Retrieval-Retro utilizes a Neural Reaction Energy (NRE) retriever trained to predict formation enthalpy using the Materials Project DFT database of approximately 80,000 computed compounds (Jain et al., 2013), but this approach does not fully exploit domain-specific data. Another limitation in extrapolation capabilities arises from the embedding design in previous approaches. Specifically, these methods embed precursor and target materials in disjoint spaces, which hinders their ability to generalize effectively.

To address these gaps, we propose **Retro-Rank-In**, a unified framework for identifying and ranking precursor sets (Figure 2). Retro-Rank-In consists of two core components: a composition-level transformer-based materials encoder, which generates chemically meaningful representations of both target materials and precursors, and a Ranker that evaluates chemical compatibility between the target material and precursor candidates. The Ranker is specifically trained to predict the likelihood that a target material and a precursor candidate can co-occur in viable synthetic routes.

Table 1: **Comparison of precursor planning methods.** ElemwiseRetro (template-based), Synthesis Similarity & Retrieval-Retro (retrieval-based), and our ranking-based approach compared for model capabilities.

Model	Discover new precursors	Chemical domain knowledge	Extrapolation to new systems
ElemwiseRetro (Kim et al., 2022)	✗	Low	Medium
Synthesis Similarity (He et al., 2023)	✗	Low	Low
Retrieval-Retro (Noh et al., 2024)	✗	Medium	Medium
Retro-Rank-In (Ours)	✓	Medium	High

Our key contributions are as follows:

- *Increased flexibility*: During inference, Retro-Rank-In enables the selection of new precursors not seen during training. This is crucial for exploring novel compounds as it allows the incorporation of a larger chemical space into the search for new synthesis recipes (McDermott et al., 2023).
- *Incorporation of broad chemical knowledge*: We leverage large-scale pretrained material embeddings to integrate implicit domain knowledge of formation enthalpies and related material properties.
- *Joint embedding space*: By training a pairwise ranking model, we embed both precursors and target materials within a unified embedding space, thereby enhancing the model's generalization capabilities.
- *Analysis of sequential models*: We compare our method against autoregressive generation approaches, demonstrating that our framework provides a more robust and accurate alternative, particularly for tasks requiring the simultaneous evaluation of multiple precursors instead of sequential modeling.

2 PRELIMINARIES

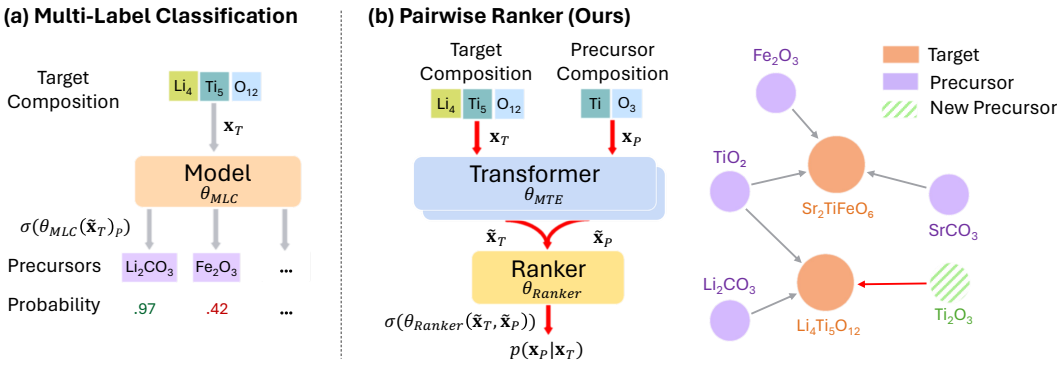

Figure 2: **Learning paradigms for inorganic retrosynthesis** (a) Multi-label classification-based approaches, which constitute current state-of-the-art models (Noh et al., 2024), inherently predict known precursors P from a fixed candidate set. (b) Our approach (Retro-Rank-In) overcomes this limitation by embedding both precursor and target materials into a shared latent space and predicting their chemical compatibility in synthetic routes. This enables extrapolation beyond known precursors, allowing the model to propose novel synthesis pathways for unseen materials. The red links highlight an exemplary case for link prediction between a target and a precursor.

Retrosynthesis. During a forward synthesis process in inorganic chemistry, a set of precursors $P_1, \ldots, P_m$ are mixed and heated to form a target material T. The inverse problem, namely retrosynthesis, devises a combination of precursors as a set that reacts to form a desired target material. Retrosynthesis involves working backward from a target compound (e.g., $Li_7La_3Zr_2O_{12}$) to deduce a set of simpler precursor compounds (e.g., $\{LiOH, La_2O_3, ZrO_2\}$) that can feasibly synthesize the desired product. However, this is an under-determined problem as there are many possible sets of valid precursors, which, under the right reaction conditions, might form the target compound. Moreover, the feasibility of a given precursor set can be quantified by considering factors such as the required synthesis pressure and temperature, the cost of precursor materials, and the yield of the process. Inspired by the fact that some precursor combinations are more advantageous than others, we formulate the learning problem as a ranking task over precursor sets.

Learning problem. Building upon previous work (He et al., 2023; Noh et al., 2024), our objective is to predict a ranked list of precursor sets, denoted as

$$(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_K). \tag{1}$$

Each precursor set $\mathbf{S} = \{P_1, P_2, \ldots, P_m\}$ consists of m individual precursor materials, where each element P_i represents a single precursor and the number m can vary for each set.

The ranking indicates the predicted likelihood of each precursor set forming the target material. Historically reported synthesis routes from the scientific literature are considered correct predictions.

While prior work focuses on learning a multi-label classifier θ_{MLC} over a predefined set of precursors/classes (He et al., 2023; Noh et al., 2024), we redefine the problem to learn a pairwise ranker θ_{Ranker} of a precursor material P conditioned on target T, see Figure 2. Thus, our reformulation of the learning problem enables inference on entirely novel precursors and precursor sets, a capability that has not been achieved by previous methods. In addition, we know that datasets in chemistry are highly imbalanced, with a large number of possible precursors and only a few positive labels. Our pairwise scoring approach allows for custom sampling strategies, including negative sampling, to improve balance.

Compositional representation. For a given target material T, we represent its elemental composition as a vector $\mathbf{x}_T = (x_1, x_2, \ldots, x_d)$, where each x_i corresponds to the fraction of element i in the compound, and d is the count of all considered chemical elements. For example, titanium dioxide (TiO_2) can be expressed as a composition vector where titanium (Ti, atomic number 22) and oxygen (O, atomic number 8) contribute respective fractions $x_{22} = \frac{1}{3}$ and $x_8 = \frac{2}{3}$.

3 RETRO-RANK-IN

3.1 EMBEDDING MODEL

We use a multi-task pretrained transformer-based encoder θ_{MTE}, adapted from (Prein et al., 2023) and (Wang et al., 2021), to map each composition $\mathbf{x}$ to an h-dimensional latent representation $\tilde{\mathbf{x}} = \theta_{\text{MTE}}(\mathbf{x})$, with $\tilde{\mathbf{x}} \in \mathbb{R}^h$.

Input sequence construction. First, each element i with a non-zero component is mapped to a learned elemental embedding $\mathbf{e}_i$. During pretraining, the model is initialized with mat2vec elemental embeddings (Murdock et al., 2020). To account for the continuous stoichiometric fraction x_i, we apply a sinusoidal fractional embedding $\mathbf{f}_i$, inspired by positional encodings in transformers (Vaswani, 2017). We then sum the elemental embeddings and their fractional encodings to form a single per-element embedding:

$$\mathbf{z}_i = \mathbf{e}_i + \mathbf{f}_i. \tag{2}$$

To achieve rich compound representations, we learn a special `[CPD]` token $\mathbf{t}$ that is prepended to the sequence to serve as a global compound-level representation. Thus, the final input sequence $\mathbf{s}$ becomes:

$$\mathbf{s} = \big[\mathbf{t}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\big], \tag{3}$$

where k is the number of distinct elements in the composition. We pass $\mathbf{s}$ through three transformer encoder blocks:

$$\tilde{\mathbf{x}} = \theta_{MTE}\big(\mathbf{s}\big), \tag{4}$$

which use self-attention to contextualize each element's embedding.

Multi-task pretraining. To encourage broader generalization, we pretrain the encoder on the Alexandria database (Schmidt et al., 2024) of over two million unique compositions using a multi-task objective, as described in appendix C.1. This includes (i) *masked element prediction*, where randomly masked elements are reconstructed, (ii) *DFT-property regression*, predicting 10 computed properties (e.g., formation enthalpy, stress tensor component, band gap) by feeding the final hidden state of `[CPD]` into regression heads, and (iii) *space-group classification*, determining the space group of $\mathbf{x}$ from the same special token output.

3.2 RANKER

We introduce a binary classifier $\mathcal{B} : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{P}$ to evaluate precursor-target pair relevance. Given a target representation $\tilde{\mathbf{x}}_T \in \mathbb{R}^h$ and precursor representation $\tilde{\mathbf{x}}_P \in \mathbb{R}^h$, we compute the probability $p \in \mathbb{P}$ of the precursor forming a valid synthetic path to the target. The training dataset consists of

balanced positive and negative pairs, effectively forming a bipartite edge prediction problem akin to recommender systems (Gao et al., 2022).

During inference, for a given target $\mathbf{x}_T$, we compute $p(\mathbf{x}_P|\mathbf{x}_T) = \theta_{\text{Ranker}}(\tilde{\mathbf{x}}_T, \tilde{\mathbf{x}}_P)$ for each precursor in the dataset. The top-K precursors, determined by probability ranking, are combined into valid sets satisfying elemental completeness and predefined cardinality constraints (Noh et al., 2024). Assuming precursor independence, we compute the joint probability of each set S as:

$$p_S = \prod_{i=1}^{m} p_i, \tag{5}$$

where m denotes the number of precursors in the set. We evaluate the final ranked precursor sets, ordered by p_S, using Top-K metrics.

3.3 Model training

We collect target–precursor pairs $(\mathbf{x}_T, \mathbf{x}_P)$ from known synthesis routes, and label them $y \in \{0, 1\}$. To balance classes, we sample positive ($y = 1$) and negative ($y = 0$) pairs with equal probability. Negative examples are generated by randomly selecting precursors that share at least one element other than oxygen with the target material but are not used together with the target in the training set. This approach ensures that the negative examples are chemically relevant. Each composition $\mathbf{x}$ is encoded by θ_{MTE}, producing $\tilde{\mathbf{x}}$. For a target–precursor pair $(\mathbf{x}_T, \mathbf{x}_P)$, we concatenate their embeddings $\left[\tilde{\mathbf{x}}_T \,\|\, \tilde{\mathbf{x}}_P\right]$ and pass them to the ranker θ_{Ranker} (Figure 2, b.) to predict the precursor probability $p(\mathbf{x}_P|\mathbf{x}_T)$. We train using binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big], \tag{6}$$

where $p_i = p(\mathbf{x}_P|\mathbf{x}_T)$ and $|\mathcal{D}|$ is the total number of samples. At test time, given a target composition $\mathbf{x}_T$, we compute $p(\mathbf{x}_P|\mathbf{x}_T)$ for each candidate precursor $\mathbf{x}_P$ and select the highest-ranked precursors for further evaluation (see appendix E).

4 Experiments

We begin by detailing our experimental setup, datasets, evaluation protocols, and baseline comparisons. We then present empirical results, followed by an in-depth analysis of our approach, including extensive ablation studies.

4.1 Setup

Datasets. We build upon the dataset introduced by (Kononova et al., 2019), which has been widely used in recent studies (Noh et al., 2024; He et al., 2023). Curated from the literature using paragraph- and phrase-level NLP classifiers, this dataset captures solid-state reactions, including byproducts, targets, and precursors, and comprises 33,343 synthesis recipes extracted from published sources.

However, this dataset has some incomplete or ambiguous entries. To ensure data quality, we apply several preprocessing steps to validate chemical formulas. Specifically, we exclude entries containing variables such as *b* and *c* or other symbolic placeholders, retaining only those with explicitly defined stoichiometries and valid element symbols from the periodic table. Additionally, we enforce a constraint requiring that all elements in the target material – except for C, O, H, and N – must also be present in the precursor materials. A viable assumption to make is that new elements cannot form during solid-state reactions.

After preprocessing, the dataset consists of 18,804 entries, of which 9,255 are unique, i.e., we exclude permutations of precursor sets. Consequently, over half of the dataset consists of duplicate entries.

Previous studies (Noh et al., 2024; He et al., 2023; Kim et al., 2022) have employed a year-based data split to construct a materials discovery setting, using data reported up to and including 2014 for training and validation, while reports after 2014 serve as the test set. However, we recognize that a high number of duplicated synthesis recipes can inflate a model's performance metric in predicting

synthesis routes for novel materials. This prevalence of duplicate entries aligns with observations in related fields, such as organic retrosynthetic planning (Bradshaw et al., 2025), where repeated recipes are frequently encountered due to recurring synthesis reports. To address these limitations, we propose augmenting the existing year-based split with two additional evaluation settings, resulting in three distinct datasets that present a more challenging setup for assessing extrapolation. More details on the datasets can be found in appendix B.

- **Complete Reaction Archive (CRA)**: Includes all entries, retaining duplicate precursor-target combinations, as conducted in previous works (Noh et al., 2024; He et al., 2023; Kim et al., 2022).
- **Distinct Reactions (DR)**: This dataset focuses exclusively on unique precursor-target combinations, represented as $\{\mathbf{x}_T, \mathbf{x}_{P_1}, \ldots, \mathbf{x}_{P_m}\}$. Duplicate entries are removed, ensuring that each reaction is represented only once. This split emphasizes the model's ability to learn distinct chemical pathways.
- **Novel Material Systems (NMS)**: Ensures that no material system – defined by the set of elements in the target material – overlaps between the training and test sets. For instance, $\mathrm{Fe}_a\mathrm{P}_b$ samples (where $a, b > 0$) appear either only in the train/validation split or in the test split. This setting enables the evaluation of the model's ability to extrapolate to entirely new systems.

Evaluation. Following recent works (Noh et al., 2024; He et al., 2023; Kim et al., 2022), we employ Top-K exact match accuracy to assess the performance of our binary classifier $\mathcal{B}$ in the inorganic retrosynthesis task. Additional details are provided in appendix E.

Let the ground-truth precursor set for a target composition $\mathbf{x}_T$ be denoted as $\mathbf{S}_{\text{true}}$, with its length defined as $m = |\mathbf{S}_{\text{true}}|$. Based on the probabilities p_i predicted by the model, we construct precursor sets as described in section 3.2. A valid set is defined by encompassing all elements of the target compound. These sets are then sorted in descending order of their joint probability. For the Top-K exact match accuracy, we select the Top-K subset of the sorted sets and compare the ground-truth precursor set $\mathbf{S}_{\text{true}}$ against each set. If a match is found, the corresponding set is assigned a score of 1; otherwise, it is scored as 0. This process is repeated for all target compositions in the test dataset, and the scores are averaged to compute the overall score. To ensure objective benchmarking, we apply the same set construction logic across all models.

Baselines. We evaluate our approach against three baseline methods for inorganic synthesis planning, as proposed in prior literature (Noh et al., 2024; He et al., 2023; Kim et al., 2022). These methods formulate precursor prediction as multi-label classification, outputting a vector of dimension N, where N represents the number of unique precursors in the dataset. He et al. (2023) introduced a masked precursor completion (MPC) task, where attention layers are employed to contextualize the representations of precursors and target materials. The model utilizes these representations to reconstruct the precursor materials. Subsequently, these learned representations are leveraged to search a knowledge base of known target materials, facilitating the transfer of precursors from known materials to novel target materials. Noh et al. (2024) expands on this by adding a neural reaction energy (NRE) retriever to the MPC task. The NRE module employs a pretrained formation enthalpy predictor to retrieve energetically favorable candidate precursors from a knowledge base. The combined model, namely Retrieval-Retro (Noh et al., 2024), then learns to use information from both modules jointly to predict precursors via a multi-label classification objective.

The approach proposed by Kim et al. (2022) utilizes a heuristic-based method integrated with a classification task through a multilayer perceptron. It extracts source elements from the input target material composition and predicts corresponding templates from a set of 60 precursor templates. The concatenation of the source element and its correspondingly predicted template forms the precursor compound.

Furthermore, we incorporate three composition-based material representation strategies to predict precursor occurrence via multi-label classification. First, we use a sparse composition approach encoding each element's fraction in a dedicated input dimension. Second, we apply CrabNet (Wang et al., 2021), a transformer encoder model designed to contextualize elemental embeddings. Finally, we test MTEncoder (see section 3.1), which uses a similar transformer-based architecture

but benefits from extensive pretraining on large-scale materials data. In all cases, we pair these representations with feedforward layers that output probabilities for each potential precursor via multi-label classification.

Table 2: **Performance comparison** Different models were evaluated across three datasets: (a) Complete Reaction Archive, (b) Distinct Reactions, and (c) Novel Material Systems. Bold values indicate the best performance and underline the second best. All scores are reported as averages over five runs, with standard deviations in parentheses.

Model	(a) Complete Reaction Archive				(b) Distinct Reactions				(c) Novel Materials Systems			
	Top-K Accuracy ↑				Top-K Accuracy ↑				Top-K Accuracy ↑			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Composition	64.72 (0.18)	77.68 (0.37)	81.17 (0.45)	83.44 (0.42)	42.31 (1.15)	55.35 (1.00)	59.36 (0.95)	62.86 (0.80)	43.57 (0.37)	55.26 (0.58)	58.27 (0.51)	60.01 (0.57)
ElemwiseRetro (Kim et al., 2022)	64.56 (0.19)	70.00 (0.12)	71.27 (0.09)	72.82 (0.06)	48.72 (1.15)	55.19 (0.32)	57.30 (0.21)	59.08 (0.22)	46.70 (1.55)	52.90 (0.05)	54.27 (0.17)	56.66 (0.23)
SynthesisSimilarity (He et al., 2023)	62.25 (0.75)	73.19 (0.55)	76.03 (0.43)	78.51 (0.30)	40.40 (0.49)	53.48 (0.31)	57.61 (0.53)	61.24 (0.54)	36.67 (0.40)	47.60 (0.80)	50.97 (1.01)	53.42 (0.94)
CrabNet (Wang et al., 2021)	<u>66.66</u> (0.43)	79.73 (0.16)	82.98 (0.37)	85.24 (0.22)	48.68 (0.51)	63.30 (0.70)	67.26 (0.58)	69.70 (0.66)	<u>48.54</u> (0.47)	59.93 (0.32)	62.44 (0.30)	63.70 (0.32)
Retrieval-Retro (Noh et al., 2024)	66.22 (0.61)	77.45 (0.27)	81.01 (0.35)	84.28 (0.43)	**49.31** (0.62)	62.70 (0.36)	67.30 (0.83)	<u>71.25</u> (0.57)	48.05 (0.58)	60.77 (1.26)	64.26 (1.18)	<u>67.68</u> (1.29)
MTEncoder	**67.04** (1.77)	<u>80.53</u> (2.47)	<u>83.89</u> (2.40)	<u>86.10</u> (1.69)	<u>49.01</u> (0.54)	<u>64.59</u> (0.29)	<u>68.78</u> (0.397)	71.24 (0.51)	**49.35** (0.40)	<u>61.98</u> (0.54)	<u>64.74</u> (0.59)	65.94 (0.45)
Retro-Rank-In (Ours)	66.55 (0.43)	**80.57** (0.90)	**85.89** (0.61)	**89.85** (0.81)	48.93 (0.50)	**65.45** (0.31)	**72.51** (0.69)	**78.48** (0.82)	47.36 (1.05)	**63.27** (1.95)	**69.92** (1.74)	**76.20** (1.86)

4.2 RESULTS

Table 2 highlights key findings that both align with and extend prior research. Consistent with Noh et al. (2024), models that explicitly capture interactions between elements (e.g., MTEncoder, CrabNet) outperform simpler composition-based baselines. Notably, MTEncoder benefits from its domain-informed pretraining, which enhances material representations by transferring knowledge from compound properties related to synthesis, such as spacegroup and formation enthalpy, leading to improved performance. All methods achieve high accuracy in the interpolation-focused CRA dataset (Table 2, a.), obtaining strong Top-1 exact match accuracy. We observe that while methods perform similarly at the Top-1 level, they do show a greater decline in accuracy compared to Retro-Rank-In at higher Top-K settings. We attribute this decline to the imbalanced multi-label training objective, which leads to a skewed probability distribution, many predictions are concentrated near 0 or 1, as shown in Figure 10, making it less suitable for the ranking task. Additionally, we see only limited benefit of the retrieval-based augmentation of target material information for methods (Noh et al., 2024; He et al., 2023).

To investigate how model performance is influenced by predictions in more challenging extrapolative settings, we curate two new datasets, one of deduplicated, distinct reactions (DR) and another of novel material systems (NMS), are introduced to evaluate model performance in this setting where no material system is shared between the training and test sets.

Distinct reactions. In the distinct reactions setting (Table 2, b.), performance declines (compared to the CRA setting) across all methods due to the exclusion of overlapping reactions between training and test data, underscoring the inherent difficulty of true extrapolation beyond memorized training examples. Notably, while the Top-1 gap narrows and Retrieval-Retro surpasses our approach in this metric, Retro-Rank-In maintains high Top-K exact match accuracy across the Top-3, Top-5, and Top-10 evaluations, widening the gap to other approaches. We hypothesize that Retro-Rank-In's advantage stems from its ability to generate a more smoothly distributed ranking of candidates, leading to better-calibrated scores for non-trivial precursor candidates. We show this in Figure 10, where Retro-Rank-In predicts a more diverse set of precursors compared to the next best method. Additionally, we find that MTEncoder achieves competitive performance, highlighting the significance of pretrained and domain-informed embeddings. However, its performance declines for larger values of K, likely due to the challenges associated with the multi-label classification setup. This comparison highlights the robustness of Retro-Rank-In and its ability to generalize effectively even when other methods struggle due to increased deduplication, suggesting that performance gains reported in previous works may have been inflated by the presence of near-duplicate entries.

Novel materials systems. However, the train-test split in the DR setting still contains compositions that are similar (e.g., $Li_5La_3Ta_2O_{12}$ in training vs. $LiLaTa_2O_7$ in testing). As such, we evaluate the models on the most difficult setting (NMS) (Table 2, c.), which shows a further widening of the train-test gap. All models face greater difficulty in this extrapolative setting, reflecting the complexity of predicting synthesis routes when training and testing data diverge more substantially. In this setting, the domain-informed approaches, Retrieval-Retro and MTEncoder, achieve strong performance for Top-1 set prediction. Notably, Retro-Rank-In performs competitively with Retrieval-Retro on the Top-1 metric.

Diversity at no cost of performance. Notably, Retro-Rank-In outperforms the next best models (MTEncoder and Retrieval-Retro) as more precursor sets (Top-K) are considered (Figure 3). Interestingly, the performance gap between the two models widens from 2.5% for $K = 3$ to 8.5% for $K = 10$. This is significant, as this shows that Retro-Rank-In is capable of generating valid precursor sets even at high K values. In contrast, this effect is less pronounced in the baseline methods. We hypothesize that this bifurcation of performance (Figure 3) at higher values of K is a result of a higher diversity of valid precursor predicted precursors as previously shown in Figure 9. Based on domain knowledge, this result is compelling, as the target-precursor relationship is inherently *one-to-many*, i.e., the same target can be synthesized with multiple possible precursor sets. As such, the high diversity of predictions is a testament to our approach to better capturing such a one-to-many relationship. From an experimental point of view, this is useful as experimentalists may prefer a diverse set of synthesis recipes instead of a few (possibly due to the availability of compounds in the lab, safety, or apparatus constraints).

Generalization to new precursors. In retrosynthesis for materials discovery, experimental material scientists start from a target compound to identify potential precursor candidates. This process involves screening extensive databases to find suitable compounds. As a case study, consider the target compound Cr_2AlB_2. A search in the Materials Project (Jain et al., 2013) database yields potential precursor compounds such as B, Cr, Al, AlB_2, $Al_{45}Cr_7$, Al_8Cr_5, $AlCr_2$, CrB, CrB_4, Cr_3B_4, Cr_5B_3, Cr_2B_3, CrB_2, Cr_2B, and Cr_4B. We train our model on a training set where none of these compounds were present, a scenario in which all previous methods would be *inapplicable*. Inputting these into our model yields the reported synthesis route of reacting $CrB + Al \rightarrow Cr_2AlB_2$ as the third highest-ranked synthesis recipe. Table 3 shows four more examples of reactions where Retro-Rank-In is able to correctly predict as its top-ranked precursor set, while Retrieval-Retro falls short (predicts $\emptyset$) due to it being limited to precursors seen during training. This underlines that Retro-Rank-In can lead to new precursor candidates, paving the way for more novel, potential synthesis pathways for target compounds.

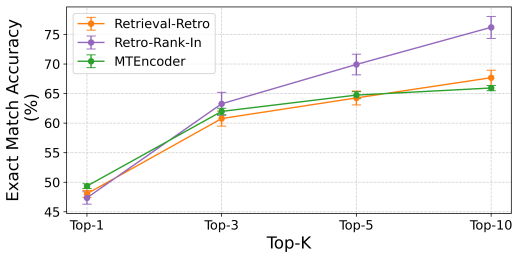

Figure 3: **Comparison of Top-K accuracy.** Comparison of Retrieval-Retro and Retro-Rank-In on the Novel Materials Systems dataset (c). We see the performance gap between both approaches widening, especially for larger K.

Suitability of sequence models. We extend our method to select precursors sequentially from a given set. Specifically, we build the precursor set S by adding one precursor at a time, allowing the model to learn conditional probabilities of co-occurrences $p(\mathbf{x}_P|\mathbf{x}_T)$. However, when evaluated on the Distinct Reaction dataset, this approach shows limited competitiveness, yielding exact match accuracies of 18.71%, 34.72%, 39.83%, and

Table 3: **Extrapolation to new precursors**: Examples from the NMS dataset where the ground truth precursors have not been part of the training set.

Target material	Ground truth	Retro-Rank-In (ours)	Retrieval-Retro
$Ba(GaSb)_2$	{GaSb, Ba}	{GaSb, Ba}	$\emptyset$
Na_5NpO_6	{Na_2CO_3, NpO_2}	{Na_2CO_3, NpO_2}	$\emptyset$
AlBMo	{BMo, Al}	{BMo, Al}	$\emptyset$
Ga_2Mo_2C	{Ga, Mo_2C}	{Ga, Mo_2C}	$\emptyset$

43.78% for Top-1, Top-3, Top-5, and Top-10 predictions, respectively. We attribute these modest results largely to the infrequency of strongly correlated precursor pairs (Figure 7), which limits the model's ability to learn meaningful joint distributions. Moreover, a larger model to process the context of precursors already assigned to the set presents significant challenges when training in the low-data regime of our task.

4.3 Model analysis

Learning a pairwise ranker. Moreover, we examine how reformulating the multi-label classification problem into learning a pairwise ranking influences the model's performance. For this analysis, we select MTEncoder embeddings. Table 4 depicts the results of those combinations. Notably, the variants that learn the pairwise ranking demonstrate a significant performance enhancement over alternatives.

Table 4: **Ablation on learning problem formulation.** We evaluate MTEncoder embeddings applied to either pairwise ranking or multi-label classification. Bold denotes the best performance.

Embedding	**Pairwise**	**Top-K Accuracy ↑**			
		Top-1	Top-3	Top-5	Top-10
MTEncoder	✗	**49.01** (0.54)	64.59 (0.29)	68.78 (0.40)	71.24 (0.51)
MTEncoder	✓	48.93 (0.50)	**65.45** (0.31)	**72.51** (0.69)	**78.48** (0.82)

5 Concluding Remarks

Limitations. While our approach represents a significant advancement, enabling the applications to novel synthesis routes and achieving state-of-the-art performance, it also has several limitations. Key synthesis parameters such as temperature, duration, and pressure, which are critical to determining the final synthesized materials, are not explicitly modeled (Huo et al., 2022). Additionally, precursor interactions can lead to the formation of intermediate compounds and byproducts, which are not captured by our current method.

Moreover, incorporating crystallographic structure data could further enhance predictive performance. However, the scarcity of datasets that integrate reaction pathways with structural information presents a challenge. Despite this, our approach is designed to be extensible and can incorporate such data when it becomes available. Resources like the Inorganic Crystal Structure Database (ICSD) (Zagorac et al., 2019) provide extensive collections of crystal structures, which could facilitate future improvements.

Summary. In this work, we introduced Retro-Rank-In, a novel ranking-based framework for inorganic retrosynthesis planning that implicitly incorporates broad chemical domain knowledge. Our approach redefines precursor prediction by learning a pairwise ranker that generalizes beyond known precursors, overcoming prior limitations and enabling the discovery of completely novel synthesis recipes. Comparative evaluations show that Retro-Rank-In sets the new state-of-the-art, particularly excelling in out-of-distribution scenarios. We will release the code upon acceptance to enable efficient synthesis planning throughout research labs.

Future work. We identify several promising directions for future research. First, integrating structural data into precursor ranking could enhance prediction accuracy by better capturing crystallographic similarities. Larger pretrained models (Liao & Smidt, 2022; Neumann et al., 2024) could further refine structural understanding, incorporating domain knowledge crucial for precursor selection. Additionally, modeling the precursor ranking as a direct ranking between a target and two precursor candidates could explicitly enable the model to choose between precursors, improving interpretability and decision-making. Finally, analyzing attention patterns and the learned chemical space could provide insights into how the model captures chemical compatibility, revealing implicit reaction rules.

REFERENCES

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

Matteo Bianchini, Jingyang Wang, Raphaële J Clément, Bin Ouyang, Penghao Xiao, Daniil Kitchaev, Tan Shi, Yaqian Zhang, Yan Wang, Haegyeom Kim, et al. The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature materials*, 19(10):1088–1095, 2020.

John Bradshaw, Anji Zhang, Babak Mahjour, David E Graff, Marwin HS Segler, and Connor W Coley. Challenging reaction prediction models to generalize to novel chemistry. *arXiv preprint arXiv:2501.06669*, 2025.

Pedro Domingos. The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999. ISSN 1384-5810. doi: 10.1023/a:1009868929893.

Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 1623–1625, 2022.

Tanjin He, Haoyan Huo, Christopher J Bartel, Zheren Wang, Kevin Cruse, and Gerbrand Ceder. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science advances*, 9(23):eadg8180, 2023.

Haoyan Huo, Christopher J Bartel, Tanjin He, Amalie Trewartha, Alexander Dunn, Bin Ouyang, Anubhav Jain, and Gerbrand Ceder. Machine-learning rationalization and prediction of solid-state synthesis conditions. *Chemistry of Materials*, 34(16):7323–7336, 2022.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024.

Christopher Karpovich, Elton Pan, Zach Jensen, and Elsa Olivetti. Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction. *Chemistry of Materials*, 35(3):1062–1079, 2023.

Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, et al. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling*, 60(3):1194–1201, 2020.

Kun Joong Kim, Moran Balaish, Masaki Wadaguchi, Lingping Kong, and Jennifer LM Rupp. Solid-state li–metal batteries: challenges and horizons of oxide and sulfide solid electrolytes and their interfaces. *Advanced Energy Materials*, 11(1):2002689, 2021.

Seongmin Kim, Juhwan Noh, Geun Ho Gu, Shuan Chen, and Yousung Jung. Element-wise formulation of inorganic retrosynthesis. In *AI for Accelerated Materials Design NeurIPS 2022 Workshop*, 2022.

Seongmin Kim, Yousung Jung, and Joshua Schrier. Large language models for inorganic synthesis predictions. *Journal of the American Chemical Society*, 146(29):19654–19659, 2024.

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, 2019.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Rubayyat Mahbub, Kevin Huang, Zach Jensen, Zachary D Hood, Jennifer LM Rupp, and Elsa A Olivetti. Text mining for processing conditions of solid-state battery electrolytes. *Electrochemistry Communications*, 121:106860, 2020.

Shreshth A Malik, Rhys EA Goodall, and Alpha A Lee. Predicting the outcomes of material syntheses with deep learning. *Chemistry of Materials*, 33(2):616–624, 2021.

Matthew J McDermott, Brennan C McBride, Corlyn E Regier, Gia Thinh Tran, Yu Chen, Adam A Corrao, Max C Gallant, Gabrielle E Kamm, Christopher J Bartel, Karena W Chapman, et al. Assessing thermodynamic selectivity of solid-state reactions for the predictive synthesis of inorganic materials. *ACS Central Science*, 9(10):1957–1975, 2023.

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, 2023.

Akira Miura, Christopher J Bartel, Yosuke Goto, Yoshikazu Mizuguchi, Chikako Moriyoshi, Yoshihiro Kuroiwa, Yongming Wang, Toshie Yaguchi, Manabu Shirai, Masanori Nagao, et al. Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis. *Advanced Materials*, 33(24):2100312, 2021.

Nicholas Monath, Manzil Zaheer, Kelsey Allen, and Andrew McCallum. Improving dual-encoder training through dynamic indexes for negative mining. In *International Conference on Artificial Intelligence and Statistics*, pp. 9308–9330. PMLR, 2023.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

Ryan J Murdock, Steven K Kauwe, Anthony Yu-Tung Wang, and Taylor D Sparks. Is domain knowledge necessary for machine learning materials properties? *Integrating Materials and Manufacturing Innovation*, 9:221–227, 2020.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

Heewoong Noh, Namkyeong Lee, Gyoung S Na, and Chanyoung Park. Retrieval-retro: Retrieval-based inorganic retrosynthesis with expert knowledge. *arXiv preprint arXiv:2410.21341*, 2024.

Thorben Prein, Elton Pan, Tom Doerr, Elsa Olivetti, and Jennifer LM Rupp. Mtencoder: A multi-task pretrained transformer encoder for materials representation learning. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*, 2023.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J Medford, and David S Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture, 2024.

Nathan J Szymanski and Christopher J Bartel. Computationally guided synthesis of battery materials. *ACS Energy Letters*, 9:2902–2911, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. npj computational materials 7 (1): 77, 2021.

Dejan Zagorac, H Müller, S Ruehl, J Zagorac, and Silke Rehme. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *Journal of applied crystallography*, 52(5):918–925, 2019.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Yuntong Zhu, Ellis R Kennedy, Bengisu Yasar, Haemin Paik, Yaqian Zhang, Zachary D Hood, Mary Scott, and Jennifer LM Rupp. Uncovering the network modifier for highly disordered amorphous li-garnet glass-ceramics. *Advanced Materials*, 36(16):2302438, 2024.

A NOTATION

Table 5: Mathematical notation overview.

Symbol	Domain	Definition
General		
$\mathbf{x} = (x_1, \ldots, x_d)$	$\mathbb{R}^d$	Material composition
x_i	$\mathbb{R}$	Stochiometric fraction of element i
$\tilde{\mathbf{x}}$	$\mathbb{R}^h$	Material embedding
P	$\mathbb{R}^d$	Precursor material
S	$\mathbb{R}^{m \times d}$	Precursor set
T	$\mathbb{R}^d$	Target material
p	$\mathbb{P}$	Probability of a precursor or set
y	$\{0, 1\}$	Whether a target-precursor pair in dataset
θ	–	Parameterized learned model
$\mathbb{B}$	–	Binary classifier
$\mathcal{L}$	–	Loss function
Dimensions		
d	$\mathbb{N}$	Dimension of composition vector
h	$\mathbb{N}$	Hidden dimension
m	$\mathbb{N}$	Number of precursors per set
n	$\mathbb{N}$	Number of unique precursors per set
N	$\mathbb{N}$	Number of unique precursor in dataset
K	$\mathbb{N}$	Top-K ranked precursor sets
MTEncoder		
$\mathbf{e}$	$\mathbb{R}^h$	Learned chemical element embedding
$\mathbf{f}$	$\mathbb{R}^h$	Sinusoidal fractional embedding
$\mathbf{z}$	$\mathbb{R}^h$	Per-element embedding
$\mathbf{t}$	$\mathbb{R}^h$	Compound embedding
$\mathbf{s}$	$\mathbb{R}^{(k+1) \times h}$	MTE input sequence
k	$\mathbb{N}$	Number distinct elements in composition

Example:

$$T \leftarrow \underbrace{\{P_1, P_2, P_3\}}_{r=3} \equiv \{\underbrace{\{A, B\}}_{m_1=2}, \{A, C, F\}, \{C, G\}\} \tag{7}$$

- $n = 5$ (explanation: there are five unique precursors $\{A, B, C, F, G\}$)
- N is a large dataset-dependent number.

B DATASET

Table 6: Dataset Statistics including Train, Validation, and Test Splits

Dataset	Train	Validation	Test
Complete Reaction Archive	9715	2430	6659
Distinct Reactions	5091	1274	2893
Novel Materials Systems	3012	753	2892

C IMPLEMENTATION DETAILS

Our code will be made available on GitHub upon publication of the manuscript.

C.1 MTENCODER

Pretraining. Figure 4 presents a schematic representation of the MTEncoder architecture, which processes material compositions using a transformer-based encoder. The input consists of element tokens (e.g., Na, Fe, O) along with a special compound token (*CPD*) that aggregates information from the constituent elements. These inputs pass through transformer encoder blocks that contextualize the elements and generate a representation of the material composition. For downstream tasks, the *CPD* token serves as the learned material representation and is fed into MLPs for property prediction.

When training on multiple material properties, each property is predicted using a dedicated MLP head attached to the transformer encoder. Simultaneously with the supervised tasks, the model learns a self-supervised denoising objective. Specifically, 30% of the element tokens are randomly masked, and the model is tasked with reconstructing the original tokens from their contextualized representations, thereby enhancing the robustness and generalizability of the learned features. Ablations on the pretraining effectivness can be found here Prein et al. (2023).

Model Configuration. MTEncoder is configured as a multi-task transformer model. It employs 3 transformer layers ($N = 3$) with a model dimension of $d_{\text{model}} = 512$ and a feed-forward dimension of 2048. The encoder uses 4 attention heads and incorporates residual neural network layers with dimensions [1024, 512] to further refine the representations. Notably, no dropout is applied, and the special *CPD* token is enabled to effectively aggregate compound-level information. For the self-supervised denoising task, 30% of the element tokens are masked. Training is conducted with a base learning rate of 5×10^{-5} for 40 epochs, using a pretraining batch size of 12. Each pretraining task is weighted equally as indicated by the sampling probabilities.

Dataset. We employ the Alexandria database for pretraining Schmidt et al. (2024). We preprocess the database by selecting a single structure per composition that exhibits the lowest formation enthalpy, based on the assumption that these structures are the most stable. An overview of the pretraining tasks is provided in Table 7.

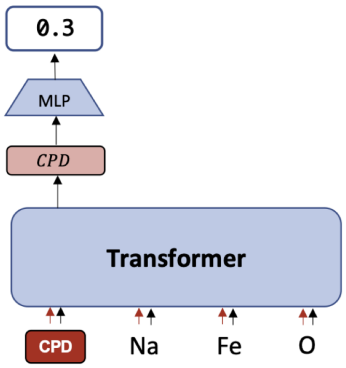

Figure 4: **MTEncoder architecture overview.** This diagram illustrates the MTEncoder framework, where material compositions are tokenized and processed through a transformer model.

C.2 TRAINING DETAILS

Model Training. We ablate over the number of layers in Retro-Rank-In to assess robustness against this hyperparameter. As shown in Figure 5, the model performs consistently well under various depths, peaking around three layers for most metrics. Top-1 accuracy dips slightly at both extremes (one and five layers) but remains stable near the center. Similar trends hold for the other metrics (Top-3, Top-5, Top-10), suggesting that while some tuning of depth may help refine results, the method is generally resilient to layer variations.

Hyperparameters For our model, we conducted further hyperparameter tuning with ranges specified in Table 8. We explored the following parameter spaces: batch size B in $\{128, 256, 512\}$,

Pretraining Tasks
Stress
Band Gap (Direct)
Band Gap (Indirect)
DOS at Fermi Level
Energy Above Hull
Formation Energy
Corrected Energy
Phase Separation Energy
Number of Sites
Total Magnetization
Space Group
Masked Element Modelling (Self-supervised)

Table 7: **Pretraining tasks** Tasks during MTEncoder pretraining, data is used from alexandria.

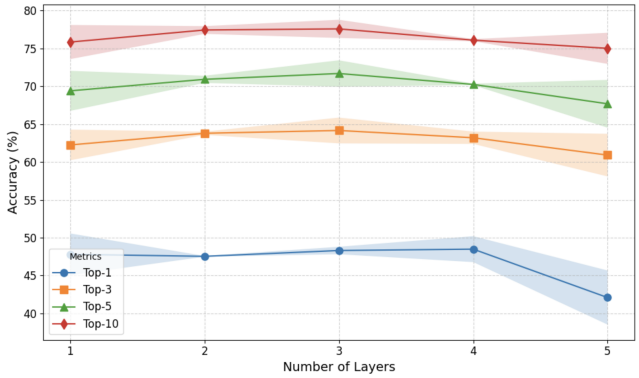

Figure 5: **Ablation for layers.** Retro-Rank-In tested for various numbers of layers. Results show the robustness of our method regarding hyperparameter choice.

number of attention heads H in {1, 4, 8}, number of feedforward (FFWD) layers L in {1,2,3,4}, learning rate η in $[10^{-5}, 10^{-3}]$ and MTEncoder learning rate η_{MT} in $[10^{-6}, 10^{-4}]$. The optimal configuration was determined based on model performance on the validation set, resulting in the following selected values: batch size of 128, 1 attention head, 3 FFWD layers, learning rate of 6.81×10^{-5} and MTEncoder learning rate of 6.37×10^{-5}. We report test performance using these optimized parameters.

Table 8: Hyperparameter configuration of Retro-Rank-In

Hyperparameters	**Configuration**	
	Search Space	Selected Values
Batch Size (B)	{128, 256, 512}	128
Attention Heads (H)	{1, 4, 8}	1
FFWD Layers (L)	{1,2,3,4}	3
Learning Rate (η)	$[10^{-5}, 10^{-3}]$	6.81×10^{-5}
MT Learning Rate (η_{MT})	$[10^{-6}, 10^{-4}]$	6.37×10^{-5}

D BASELINE METHODS

D.1 RETRIEVAL-RETRO

Retrieval-Retro Noh et al. (2024) proposes a two-stage approach to inorganic retrosynthesis that implicitly extracts the precursor information of reference materials. First, for each target material, reference materials from the knowledge base of previously synthesized materials are elaborately retrieved by two complementary models: Inspired by Synthesis Similarity He et al. (2023), the MPC retriever, trained for Masked Precursor Completion (MPC), selects reference materials sharing similar precursors. The Neural Reaction Energy (NRE) Retriever integrates domain knowledge and leverages the thermodynamic relationships between materials to identify precursor sets with a high probability of synthesizing the target. Representing target and reference materials as fully connected composition graphs, the final Retrieval-Retro stage then employs self-attention and cross-attention mechanisms to implicitly extract relevant precursor information from the reference materials and predict precursor sets based on the probability for each individual precursor.

D.2 ELEMWISERETRO

ElemwiseRetro Kim et al. (2022) proposes a template-based approach to inorganic retrosynthesis that represents target materials as fully connected composition graphs. To guide the retrosynthesis process, the researchers distinguish between two types of elements: "source elements," provided as precursors, and "non-source elements," which either appear or disappear during the reaction. For each source element in a target composition, their model predicts the most likely anionic framework—a composition of non-source elements—from a predefined set of templates. The selected source element and its template are then concatenated to form the actual precursor compound, which may be reformulated using a stoichiometric lookup table to ensure frequent and chemically valid compositions.

D.3 SYNTHESIS SIMILARITY

He et al. (2023) uses a similarity-based approach to identify precursor sets for inorganic retrosynthesis. They introduce a vector representation for Masked Precursor Completion (MPC) and chemical composition recovery tasks and use this encoding to retrieve reference materials similar to a given target. They initialize the prediction with the precursor set of the reference material and use the MPC network to complement the prediction for a valid precursor set.

D.4 MISTRAL 7B

We utilize the Mistral 7B model (Jiang, 2024) for LM-based precursor prediction. Initially, few-shot prompting was attempted, but its performance proved suboptimal. Consequently, we adopted a more structured prompting strategy. The prompt was as follows:

> "You are tasked with identifying precursors for synthesizing the target material $Mn_{0.71}Zn_{0.21}Fe_{2.08}O_4$. You should choose exactly 3 precursors. Generate 20 possible precursor material combinations in descending order of probability. Each route should represent a unique combination of precursors likely to result in the target material. Use the chemical formulas for all precursors instead of their common names. Output them in a Python list format, where each precursor is a string, and each possible combination is a list. Each list should have a length of 3. The response should only include a list of lists where the smaller the index of the list, the higher the probability that the precursor combination has. Do not add any other sentences to the response, only print the list."

This structured prompt significantly enhanced prediction accuracy, and post-processing steps such as element validation and duplicate removal further refined the results (see Table 9). However, when compared to custom expert models specifically designed for precursor prediction, Mistral 7B's performance remains lower. We attribute this discrepancy to several factors: First, as a 7-billion-parameter model, Mistral 7B has limited capacity to capture the intricate chemical knowledge and nuanced relationships that specialized models, trained on domain-specific datasets, possess. Second,

unlike expert systems that select precursors from a predefined candidate set, Mistral 7B is required to generate precursor combinations in free-text form, increasing the risk of hallucinations and variability. Third, its general-purpose training does not fully incorporate the strict chemical constraints and synthesis rules that custom expert models are optimized for. Future work could explore integrating domain-specific data and hybrid retrieval-generation approaches to better harness the efficiency of large language models for inorganic retrosynthesis.

Table 9: **Performance results for Mistral.** Mistral evaluated across three datasets: (a) Complete Reaction Archive, (b) Distinct Reactions, and (c) Novel Material Systems. Bold values indicate the best performance and underline the second best. All scores are reported as averages over five runs, with standard deviations in parentheses.

Model	(a) Complete Reaction Archive				(b) Distinct Reactions				(c) Novel Materials Systems			
	Top-K Accuracy ↑				Top-K Accuracy ↑				Top-K Accuracy ↑			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Mistral 7B	24.75	35.34	37.62	39.78	16.91	22.30	24.27	25.93	16.91	22.30	24.27	25.93

E Evaluation Protocol

To mitigate the combinatorial explosion, we first select the 30 precursor candidates with the highest probabilities from the model's predictions. Table 10 illustrates how increasing this sample size improves Top-K match accuracy but also amplifies computational complexity, as more precursor combinations must be evaluated. These selected precursors are then combined to form candidate sets, and a set is deemed valid if the union of its elements contains all elements of the target composition. Each valid set is assigned a joint probability derived from the probabilities of its individual precursors, and the valid sets are subsequently ranked in descending order by this joint probability. For the Top-K match accuracy, we focus on the subset of K highest-ranked valid sets. If the ground-truth precursor set S_{true} matches any of these K sets, we assign a score of 1 for that target; otherwise, we assign 0. Finally, we compute the overall score by averaging these individual scores across the entire test set.

We evaluate all approaches in this way, except for Mistral and ElemwiseRetro, which already output the precursor sets. Therefore, we skip the step of constructing the sets from the candidates. The remaining part of the evaluation stays the same.

Table 10: **Top-K Match Accuracy across candidate precursor sample sizes n.** Investigating the effect of varying amounts of top n candidate precursors sampled based on highest probabilities (see appendix E)), on example Top-K accuracy of Retro-Rank-In, and on the total number of evaluated precursor combinations N.

n	Top-K Accuracy ↑				N
	Top-1	Top-3	Top-5	Top-10	
10	43.22	57.05	62.34	67.95	357,416
20	43.26	57.26	62.59	68.78	7,333,299
30	43.30	57.43	62.90	69.29	58,386,823
40	43.57	57.81	63.31	69.81	320,996,717

F Further Analysis

F.1 Choice of information contextualization.

From Table 12, all of the ablation variants (self-attention, concatenation, addpooling, transformer, and meanpooling) perform within a fairly narrow range—the Top-1 through Top-10 accuracy numbers are all close, and the standard deviations also suggest there the absence of a clear superior strategy. Because the performance differences are small, the simplest method becomes the most appealing choice in practice (Occam's razor (Domingos, 1999)), as they provide results on par with the more complex approaches while being easier to implement and faster to train.

Table 11: **Ablation for ranker architecure.** We compare different ranker architectures.

Convolution	Top-K Accuracy ↑			
	Top-1	Top-3	Top-5	Top-10
Self-Attention	47.30	62.48	69.56	<u>75.99</u>
	(0.79)	(1.01)	(1.20)	(1.10)
Concatenation	47.04	62.64	**70.05**	**76.61**
	(1.76)	(1.73)	(2.13)	(1.66)
Addpooling	**48.48**	**63.36**	<u>69.96</u>	75.85
	(1.48)	(0.56)	(0.44)	(0.58)
Transformer	45.95	62.07	68.55	74.65
	(1.92)	(1.56)	(1.42)	(1.42)
Meanpooling	<u>47.78</u>	<u>63.13</u>	69.35	75.78
	(1.06)	(1.02)	(1.28)	(1.55)

F.2 Choice of ranker architectures.

As shown in Table 12, all variants (self-attention, concatenation, addpool, transformer, meanpool) perform similarly, with close Top-1 to Top-10 accuracy and overlapping standard deviations. Given these minimal differences, the simplest method is preferable for efficiency and ease of implementation (Occam's razor (Domingos, 1999)).

F.3 Hard negative mining.

In our experiments, we integrate hard negative mining into the training process to evaluate its impact on model performance. At each epoch, we increase the sampling probabilities of negative samples with high cosine similarity to the ground truth precursor set. Contrary to findings in other domains where hard negative mining enhances model accuracy Moreira et al. (2024); Monath et al. (2023), our results indicated that this technique did not improve performance in our specific application.

F.4 Hyperparameter ablation

We examined the impact of network depth on our model's performance by training Retro-Rank-In with feedforward layers ranging from 1 to 5. With each additional layer, we reduced the dimensionality by a factor of two. Our findings indicate that the model is robust to these hyperparameter variations, with a three-layer architecture yielding the highest Top-1, Top-3, and Top-5 accuracy scores. Consequently, we selected this as our default configuration (appendix C.2).

F.5 Pretraining encoder ablation

Table 12 further examines the impact of no pretraining versus pretraining for the encoder model. We observe the Top-K accuracy drastically decrease without having a pretrained encoder.

F.6 Performance across different chemistries

In Figure 6, we illustrate the correlation between the target embeddings and the ranks assigned by our model, namely the Retro-Rank-In. To achieve this, we first process the chemical composition of each target material using the `Composition` class from the `pymatgen` library, which parses the input material string and standardizes its representation to ensure consistency in the interpretation of the chemical formula. The standardized composition was then encoded using the MTEncoder model Prein et al. (2023), which maps the material string to a h-dimensional tensor representation. In this study, an embedding dimension of $h = 512$ was used to capture the essential features of each material for further computational analysis. After acquiring the MTE embeddings, we projected them into a 2D space using Principal Component Analysis (PCA), where each point's color represents its assigned rank. Warmer hues (red) represent higher ranks, while cooler tones (blue to white) indicate lower ranks. The predominance of red points on this plane indicates that most embeddings were correctly classified as Rank 1, demonstrating strong model performance. In contrast, lower-ranking

Table 12: **Ablation pretraining.** Investigating the impact of pretraining for the encoder with the Top-K accuracy.

Pretrained Encoder	Top-K Accuracy ↑			
	Top-1	Top-3	Top-5	Top-10
✗	33.24 (8.20)	53.13 (5.52)	62.70 (5.45)	71.22 (2.98)
✓	**47.04** (1.76)	**62.64** (1.73)	**70.05** (2.13)	**76.61** (1.66)

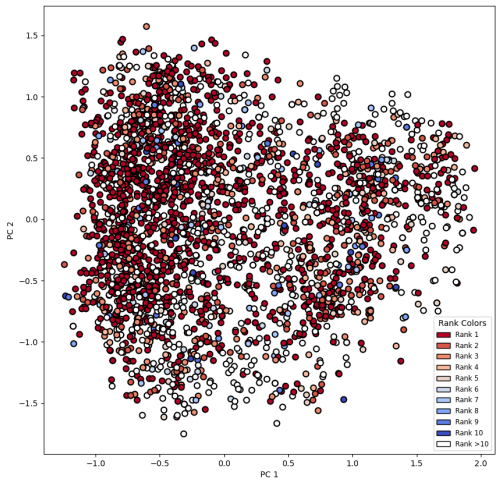

Figure 6: **PCA Visualization of target embeddings with rank-based coloring.** PCA of target embeddings, where each point is color-coded based on rank. Higher-ranked points are shown in warmer tones, while lower-ranked ones appear in cooler shades, illustrating the distribution of rankings in the embedding space.

embeddings (e.g., those in blue) appear sparser, occupying smaller regions of the plot. This suggests that while the majority of materials achieve Rank 1, fewer are assigned to lower ranks. In summary, this two-dimensional projection thus highlights the distribution of performance across the embedding space, revealing that the majority of embeddings cluster at higher ranks while relatively few reside in the lower-rank region.

F.7 PRECURSOR CORRELATION

The top plot of Figure 7 is a significant degree of positive and negative correlation for a fair amount of precursor pairs. The frequency of occurrence for the same pairs is visualized in the bottom figure. While a few unique precursor pairs show strong positive or negative correlations, the vast majority of frequently occurring pairs exhibit little to no correlation.

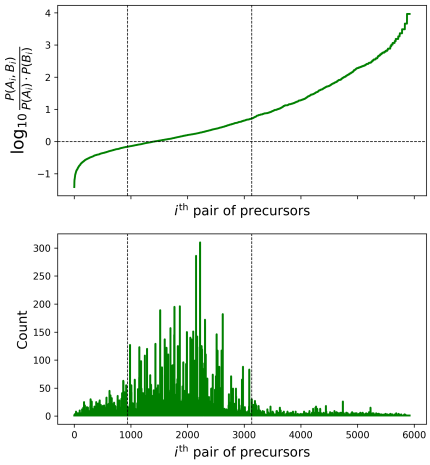

Figure 7: **Precursor pair correlation.** The plot illustrates the correlation between pairs of precursors. Each point corresponds to a unique precursor pair, sorted along the x-axis by the strength of their correlation. Correlation is quantified here by the logarithm of the ratio of their joint probability to the product of their individual probabilities.

F.8 PREDICTION DIVERSITY

Figure 8 compares the number of unique precursor combinations identified by the two models, RetrievalRetro and Retro-Rank-In, at Top-K thresholds of K=50 and K=100. We observe that

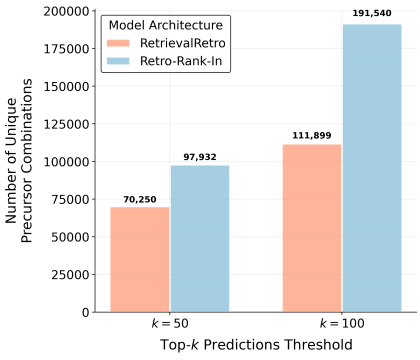

Figure 8: **Precursor set diversity.** Comparison of the number of unique precursor combinations generated by Retrieval-Retro and Retro-Rank-In.

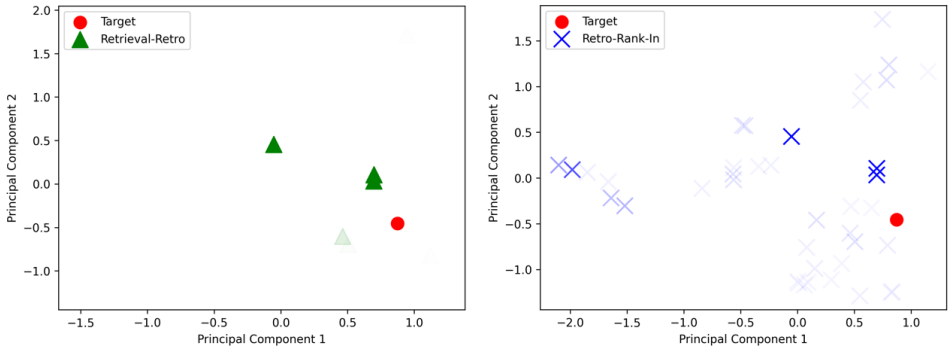

Figure 9: **Retro-Rank-In achieves higher diversity of predicted precursors.** PCA plot of a MTEncoder-encoded target material (red color) and precursors predicted by Retrieval-Retro (green triangles, left) and Retro-Rank-In (blue crosses, right). The intensity/alpha of each point is proportional to the probability assigned by each model. Clearly, Retrieval-Retro assigns high probabilities to a small number of precursors, leading to low diversity. In contrast, Retro-Rank-In assigns significant probabilities to a higher number of precursors, leading to higher diversity. Importantly, this improvement in diversity does not come at the expense of accuracy, as shown in Table 2.

Retro-Rank-In consistently generates a greater number of unique precursor combinations than RetrievalRetro. This result suggests that Retro-Rank-In can capture a broader space of potential synthetic routes, which is advantageous for identifying novel and efficient strategies in retrosynthetic planning.

This is further supported by the findings of Figure 9, which shows how each model allocates probability mass across its proposed precursor sets. In particular, RetrievalRetro tends to concentrate most of its probability on just a few highly ranked precursor combinations, as evidenced by a steep probability drop-off after its top-ranked suggestions. By contrast, Retro-Rank-In spreads its probability more evenly across a larger set of potential combinations, indicating a more diverse exploration of the synthetic space. Crucially, this broader coverage means Retro-Rank-In is less likely to overlook innovative or less obvious routes during retrosynthetic planning, offering a more comprehensive foundation for subsequent experimental validation.

Lastly, Figure 10 compares the distribution of predicted precursor probabilities (the top 60 highest values) for Retro-Rank-In (top) and Retrieval-Retro (bottom) on the Distinct Reactions test set. Retro-Rank-In produces a wider spread of mid-to-high probabilities, suggesting more nuanced confidence estimates, while Retrieval-Retro's predictions cluster near zero or one. This pattern indicates that Retro-Rank-In is better calibrated, resulting in stronger performance at higher values of K in Top-K exact match accuracy.

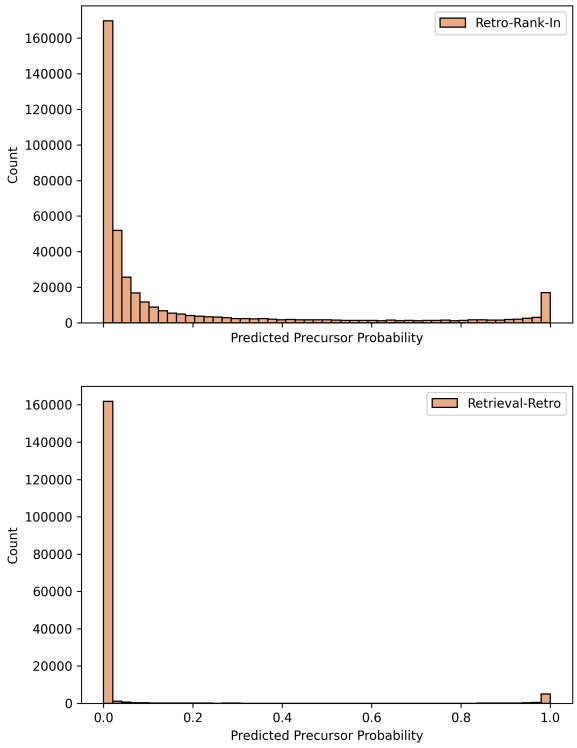

Figure 10: **Distribution of predicted probabilities.** A comparison of the top 60 highest predicted precursor probabilities across the test set of the Distinct Reactions dataset. Our approach demonstrates improved probability calibration compared to the previous state-of-the-art (Noh et al., 2024). We attribute this improvement to the class-balanced learning of a pairwise ranker, which translates to enhanced throughout performance at higher values of K in Top-K exact match accuracy.