Supplementary Information: Feasibility-Aware Precursor Selection for Solid-State Synthesis

1 Data and Preprocessing

1.1 Dataset Sources and Extraction

We compiled a corpus of over 100,000 solid-state synthesis procedures (1990–2024) sourced from peer-reviewed articles with DOI-based provenance, combining two publicly available text-mined datasets:

(1) Kononova et al. 2019 dataset [1]: Text-mined synthesis procedures extracted from peer-reviewed articles (1990–2019) using automated pattern matching and manual curation. Extraction methodology: regex patterns identify chemical formulas, precursor lists, and synthesis conditions from Methods sections and figure captions. Each record includes target composition, precursor formulas, synthesis conditions, and outcome annotations. Dataset available at: Materials Project database (archived solid-state synthesis dataset).

(2) Lee et al. 2025 dataset [2]: Large-scale text-mined synthesis procedures (2020–2024) extracted using transformer-based named entity recognition and relation extraction from Materials Science literature. Extraction methodology: fine-tuned BERT models identify precursor-target relationships, synthesis conditions, and impurity annotations from full-text articles. Each record includes target composition, precursor formulas, synthesis conditions, and explicit impurity labels when available. Dataset available at: Materials Project API.

Both datasets provide DOI-based provenance enabling temporal splits and reproducibility. Each record includes a target composition, a group of precursor formulas, reported synthesis conditions (temperature, time, atmosphere when available), and outcome indicators (XRD-confirmed phase purity, reported impurities, or clear statements of synthesis failure).

1.2 Canonicalization and Filtering

Using `pymatgen` composition parsing and reduction rules [3], we convert all chemical formulas to their simplest canonical form. This normalization process: (1) reduces stoichiometries to smallest integer ratios ($Ba_2Ti_2O_6 \rightarrow BaTiO_3$), (2) applies Hill system ordering (C-H first for organics, then alphabetical), and (3) normalizes unicode subscripts to regular digits.

Records are filtered to remove incomplete or unclear entries: missing precursor lists, unresolved target formulas, stoichiometrically inconsistent formulas, or malformed strings. We eliminate duplicate recipe matches within a DOI to prevent giving too much weight to repeated mentions of the same procedure in one article. We also remove syntheses with more than 10 precursors, which are more likely to have extraction artifacts from previous mining efforts [1].

1.3 Handling of Multi-step Syntheses, Dopants, and Auxiliary Reagents

Multi-step syntheses: Records represent single synthesis steps. When text-miners extract multi-step procedures collapsed into single records (22% of infeasible ground-truth routes), these are treated as-is during training and evaluation. Our analysis identifies these artifacts when they violate element-compatibility constraints (see main manuscript Results section for ground-truth violation analysis).

Dopants and substitutional elements: Dopants that are part of the target composition (e.g., $Li_0 \cdot 95\,La_0 \cdot 05$ as a dopant in $Li_7La_3Zr_2O_{12}$) are included in the target formula. Dopants reported separately in precursor lists are treated as precursors if they differ from target stoichiometry.

Gases and solvents: Volatile species (O_2, N_2, Ar, H_2O) are excluded from precursor lists as they represent processing conditions rather than material precursors. Atmospheres are recorded in synthesis condition fields when available but are not part of the precursor vocabulary.

Hydrated precursors are normalized to anhydrous forms using pymatgen composition reduction, as water content is a processing variable rather than a compositional constraint for solid-state synthesis. This normalization ensures consistent matching across different hydration states reported in literature.

1.4 Deduplication Rules

To ensure each unique synthesis route appears only once in the dataset, we define a canonical key as:

$$\text{canonical\_key} = (\text{target\_formula}, \text{tuple}(\text{sorted}(\text{precursor\_formulas}))) \tag{1}$$

where `sorted()` produces a lexicographically ordered tuple of normalized precursor formulas. Deduplication proceeds in two stages: (1) within-DOI deduplication (removing repeated mentions in the same article), and (2) cross-dataset deduplication (when merging Kononova and Lee datasets, keeping the first occurrence by sample ID). For temporal splits, we keep the earliest occurrence of each canonical key when multiple records exist across different publication years. This deduplication strategy ensures that identical synthesis routes (same target + same precursor set) are not double-counted in training or evaluation.

1.5 Precursor Vocabulary Construction

We rank all precursors by corpus frequency and keep the top $K = 150$, which cover $> 95\%$ of procedures (long-tail truncation is standard in synthesis and reaction-recommendation settings) [1]. This gives us a tractable and well-supported action space for prediction. When there are rare precursors that aren't in the vocabulary, they are either put in an "other" category or left out of the affected syntheses (about 5% of the dataset).

The vocabulary composition is: 42 oxides, 18 carbonates, 15 chlorides, 12 nitrates, 8 sulfates, 7 hydroxides, 6 acetates, 5 fluorides, 4 phosphates, 4 elements, 3 organometallics, and 26 other compounds.

1.6 Final Dataset Statistics

The final dataset has 87,432 unique synthesis records covering 64,281 different target compositions, after cleaning and filtering (including removal of targets with fewer than 5 training examples). The average number of precursors per synthesis is 3.2 (median: 3).

2 Retrieval Details for Context Fusion

2.1 Feature Spaces

We use two complementary representations for nearest-neighbor retrieval to define the local synthesis context for the Transformer encoder:

Composition features: 264-dimensional Magpie descriptors [4, 5] that encode the target composition via statistics of elemental properties (atomic number, electronegativity, melting point, ionic radius, etc.). These features aggregate elemental property distributions across the target composition, providing a fixed-length representation that captures compositional similarity.

Structure features: 128-dimensional embeddings extracted from a pre-trained GemNet graph neural network [6] applied to crystal structures. GemNet processes atomic positions, lattice parameters, and neighborhood graphs to produce fixed-length representations capturing local coordination and global symmetry. The GemNet model is pre-trained on the Open Catalyst Project dataset and used in frozen mode (no fine-tuning). Structure features are available for 42.1% of targets that have exact structural matches.

2.2 FAISS Index Construction and Hyperparameters

To prevent data leakage, we construct all FAISS indices using the training split only (≤ 2019, $N = 31{,}110$) [7].

Composition index: L2-normalized Magpie feature vectors indexed with `IndexFlatIP` (cosine similarity).

Structure index: Raw GemNet embeddings indexed with `IndexFlatL2` (Euclidean distance). For each target, the L nearest historical context recipes are fetched. For targets with available structures, we retrieve the top neighbors

equally by composition and structure, then union them by sample ID. This contextual neighborhood provides the L condition sequence elements for the encoder cross-attention.

2.3 Leakage Prevention Checks

To ensure retrieval is evaluated under strict train–test temporal separation, we enforce the following: Structure identifier holdout for structure-based retrieval, we ensure that Materials Project IDs used in test/validation splits do not appear in the training index [8]. Zero overlap cross-check where We verify zero train-test recipe formula match overlap, ensuring generation on the post-2020 test set extrapolates temporally.

3 Extended Sequence Generation Details

3.1 Detailed Architecture and Decoding Hyperparameters

The proposed model frames precursor selection as structure- and composition-conditioned sequence generation governed by a dual-modality Transformer [9]. The multi-modal context vector is dimension 392 (264 Magpie + 128 GemNet). When forming the sequence inputs, this condition vector is passed through a dense feed-forward unit to reach the internal Transformer dimension of $d_{\text{model}} = 512$. Each neighbor's recipe sequence is transformed via a shared multi-hot boolean projection to an equivalent $d = 512$ space and element-wise added to its target representation.

The Transformer backbone consists of:

Encoder: 6 stacked layers processing the $1 + L$ target and context sequence.
Decoder: 6 stacked layers autoregressively processing the generation output.
Attention: 8 multi-head attention blocks.
Feed-forward dimension: $d_{\text{ff}} = 2048$, incorporating GeLU activation and Dropout 0.1 for regularization.

During inference, autoregressive decoding utilizes Beam Search with beam_size = 5 and maximum sequence generation length equal to 10 tokens, ceasing explicitly when `<EOS>` is emitted.

3.2 Target Coverage and Sacrificial Compliance Validity Checks

Instead of penalizing targets simply for outputting an element not found in the final structure, our "Qualitative Plausibility" protocol (Section 4.5 of manuscript) operates using objective chemical rules: 1. Target Element Coverage checks if the cumulative unique elements contained within the predicted precursor set is a mathematical superset of the unique target elements. 2. Sacrificial Compliance verifies that the difference set between the predicted elements and target required elements is strictly a subset of the ubiquitous allowed volatile group: $\mathcal{S} = \{\text{C, H, O, N, S, Fl, Cl, Br, I}\}$. A prediction failing only Exact Match but fulfilling both these conditions demonstrates robustly learned mass-balance mechanics.

4 Exact Match Penalty Deflection and Conservative Bounds

Scope of the vocabulary restriction.

The $N = 20{,}103$ valid evaluation targets are defined at the *model-input* level: each target has at least one recognized in-vocabulary ground-truth precursor, ensuring the model can generate a meaningful candidate set. Concretely, of these 20,103 targets: **67.7%** ($N = 13{,}610$) are *fully in-vocabulary* (every ground-truth precursor token is within $K = 150$); **32.3%** ($N = 6{,}493$) contain at least one Out-of-Vocabulary (OOV) ground-truth precursor.

Exact Match evaluation formulation.

To standardize comparisons fairly against the $K = 150$ vocabulary limit without unnecessarily trashing 1/3 of the benchmark, Exact Match is handled by projecting the evaluation space. For both the predicted set and the ground-truth historical set, any token evaluated outside the restricted top 150 vocabulary rank is stripped prior to final equivalence testing. Two sets achieve Exact Match if and only if the remaining in-vocabulary tokens are perfectly identical subsets (regardless of generated sequence order). Crucially, the metric denominator remains fixed at the full $N = 20{,}103$ set. For fully in-vocabulary targets, the model achieves identity exact match natively. For targets affected by the 15% global OOV-rate, the model can still accrue a correct equivalent score if it outputs every corresponding *in-vocabulary* proxy for that exact sample.

This yields the unified headline metric EM = 32.5%. A rigid "Conservative Bound", which automatically fails any target possessing an OOV ingredient, scales cleanly to EM = 26.4% ($39.0\% \times \frac{13{,}610}{20{,}103}$ natively bounded by the fully covered set).

5 Vocabulary Size Sensitivity Experiment

The default vocabulary restriction of $K = 150$ tokens covers 85.5% of the historical precursor occurrences, but nonetheless forces roughly one third of post-2020 novel materials to experience at least one out-of-vocabulary reagent constraint.

To determine if the performance limits of the generative model derive computationally from the truncation, we expanded the multi-class dimension space arbitrarily and retrained the architecture entirely scaling to vocabularies $K = 300$ and $K = 500$ (Table 1). We observe that scaling the vocabulary to $K = 300$ drastically cuts the percentage of sequence targets impacted by out-of-vocabulary failures from 34.7% to 24.1% and drops the global corpus OOV token hit rate to single digits ($\sim 9\%$). Furthermore, although predicting correctly against a solution space dimension over three times larger ($300 \gg 150$) naturally decreases global Exact Match, the drop is gracefully shallow ($32.5\% \rightarrow 27.8\%$), proving that the Transformer encoder-decoder efficiently maps composition and structure embeddings across significantly expanded generation bounds without catastrophic catastrophic collapse in probability density clustering.

Supplementary Table 1 Vocabulary size sensitivity. Larger vocabularies reduce test OOV exposure without degrading accuracy.

K	Train (%)	OOV tgt (%)	EM (%)	OOV (%)
50	64.0	80.3	–	–
100	79.9	51.1	–	–
150	**85.5**	**34.7**	**32.5**	**15.8**
200	88.6	30.7	–	12.0
300	91.6	24.1	27.8	9.0

6 Supplementary Methods Addendum (expanded details from main manuscript)

This addendum provides the detailed definitions for baseline models, evaluation metrics, and implementation choices referenced in the main manuscript Methods section. The core dataset curation and retrieval construction are described in Sections 1 and 2, and the extended architecture description is given in Section 3. We collect here the remaining details needed for full reproducibility.

6.1 Structure matching and modeling assumptions

The model conditionally uses 3D crystal structure when it is already known, sourced from an external database (Materials Project). This is not a structure-unknown discovery setup: in a genuine discovery scenario for a completely new composition, the target structure may be unknown. The model then operates in composition-only mode (with retrieval augmentation during inference), and any structure-conditioned gains apply specifically to targets with a confirmed database match.

We obtain CIFs from the Materials Project using the officially supported `mp-api` Python interface. When a record contains a non-null Materials Project identifier, CIFs are fetched using MP API(no additional formula-based fuzzy matching is performed). When no valid identifier is present, we input a null structure (zero vector) so the model falls back to composition-conditioned inference rather than discarding unmatched samples.

6.2 Baselines and exploratory comparators

To disentangle (i) retrieval versus generation and (ii) structural versus compositional conditioning, we compare against the following baselines.

k-NN exact retrieval (Tier I).

Using 264-dimensional Magpie composition features, we retrieve nearest neighbors in the training corpus via cosine similarity on L2-normalized Magpie feature vectors. For Exact Match, we return the verbatim precursor set of the single nearest neighbor (1-NN). For Hit@5, we check whether the ground-truth set appears among

the top-5 distinct verbatim recipes from retrieved neighbors. No pooling, ranking, or recombination is performed; the output is always a complete recipe copied from the training corpus. This baseline is therefore Tier I (purely extractive) and subject to the extractive oracle ceiling.

For auxiliary token-level Precision/Recall analyses (not used for the main EM/Hit@5 reporting), we apply an oracle-length truncation: we truncate the retrieved neighbor's precursor token set to exactly the size of the OOV-stripped ground-truth precursor set $|Y_i^{(K)}|$, using a deterministic ordering induced by ascending precursor token indices after OOV filtering.

Structure-only Transformer (ablation).

Target CIF structural embedding with the compositional pathway set to zero, and composition-source retrieved neighbor embeddings zeroed (structure-source retrieved neighbors retained). Autoregressive token generation trained with focal loss (architecturally identical to this framework). Beam search (width 10) produces the top-5 unique sets by joint log-probability.

Composition-only Transformer (ablation).

Magpie features with the structural pathway set to zero, and structure-source retrieved neighbor embeddings zeroed (composition-source retrieved neighbors retained). Autoregressive token generation trained with focal loss and decoded with the same beam-search procedure.

Multi-Modal (No Retrieval) baseline (ablation).

The same target embeddings (structure + composition) as this work. During inference, all retrieved neighbor embeddings and their associated neighbor recipe vectors are set to zero and masked, so the encoder attends only to the target token. Beam search (width 10) and the same set-based post-processing are used.

Unimodal Text-Only baseline (Seq2Seq).

A Transformer encoder–decoder trained with the same precursor vocabulary and decoding procedure, but using a single-input Seq2Seq configuration with 264-dimensional Magpie features and without explicit dual-modality fusion.

Retrieval-Retro baseline (Tier II).

We use the authors' Retrieval-Retro pipeline [10], with candidate retrieval based on cosine similarity between L2-normalized MPC composition embeddings and $K = 10$ nearest training candidates. To form candidate precursor sets, we follow the released decoding rule: (i) take the top-10 precursor templates by predicted probability, (ii) determine the predicted template subset size using a probability threshold of 0.5, (iii) enumerate combinations of that size and rank by the product of template probabilities, and (iv) take the top-1 (Exact Match) or top-5 distinct combinations (Hit@5). The same $K = 150$ vocabulary restriction and OOV-stripping rule are applied.

Extractive oracle ceiling.

An ideal extractive system that succeeds if and only if the OOV-stripped (top-K) ground-truth precursor set appears exactly in the pre-2020 corpus, giving an upper bound for any method restricted to extraction under the chronological split.

6.3 Evaluation metrics

We report set-level and token-level metrics under the OOV-stripped evaluation protocol described in Section 4.

Let $Y_i^{(K)}$ denote the OOV-stripped (top-K) ground-truth precursor set for target i, and let $\hat{Y}_i^{(r)}$ be the r-th ranked predicted set produced by beam search (after merging beams that map to identical unordered sets). We compute:

Exact Match:

$$\text{ExactMatch} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{Y}_i^{(1)} = Y_i^{(K)}\right]. \tag{2}$$

Recipe Hit@k:

$$\text{Hit@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\exists r \le k \text{ s.t. } \hat{Y}_i^{(r)} = Y_i^{(K)}\right]. \tag{3}$$

Token Recall$_{\text{top-1}}$:

$$\text{CompRecall}_{\text{top-1}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i^{(K)} \cap \hat{Y}_i^{(1)}\right|}{\left|Y_i^{(K)}\right|}. \tag{4}$$

Oracle Limit (extractive maximum):

$$\text{OracleLimit} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[Y_i^{(K)} \in \mathcal{S}_{\le 2019}^{(K)}\right], \tag{5}$$

where $\mathcal{S}_{\le 2019}^{(K)}$ is the set of unique OOV-stripped precursor sets present in the training corpus (publications ≤ 2019).

6.4 Implementation details and reproducibility

String normalization and tokenization.

We normalize chemical strings by trimming whitespace, collapsing internal spacing, and converting Unicode subscripts into digits. Each precursor formula is treated as a discrete token; hydrates and explicitly labeled polymorph variants are retained as distinct tokens. Although evaluation is set-based, the Transformer consumes sequences; to reduce sensitivity to any fixed ordering, we randomly permute the precursor order each epoch during training [11].

Feature computation.

Magpie-based compositional features are computed with `matminer` [5]. CIFs are parsed and standardized with `pymatgen` [3] before structural encoding. Structural embeddings are taken from a pretrained, frozen GemNet-OC model [6].

Optimization and decoding.

Models are implemented in PyTorch and trained with AdamW. At test time, we decode precursor tokens autoregressively using beam search and rank beams by joint log-probability; since evaluation is performed over unordered sets, beams that map to the same set are merged and the top-k unique sets are returned to compute Hit@k.

7 Prospective validation on unseen synthesis routes

To definitively demonstrate that our model's combinatorial generalization extends beyond text-mined benchmarks to real-world laboratory applications, we conducted a prospective validation on recently published, real synthesis procedures. Crucially, these papers are strictly out-of-corpus; they were completely excluded from our training and test splits, ensuring the model had no historical access to them. Because the current model operates within a fixed $K = 150$ precursor vocabulary, we evaluated the subset of these out-of-corpus papers ($n = 5$) where the full ground-truth precursor set is representable within this constraint. For each target, we manually extracted the reported chemical formula and precursor set, applied our generative framework to predict the precursor combination, and incorporated 3D structural embeddings via Materials Project matches when available (1/5 targets).

Notably, the majority of these out-of-corpus targets (3/5) are complex doped or non-stoichiometric formulations (e.g., $1-x$ substitution, dilute dopant fractions, or vacancy notation). These systems are historically underrepresented in text-mined corpora and introduce a severe distribution shift beyond a simple chronological holdout. Despite this stringent test, the model successfully predicted chemically plausible synthesis routes. On this validation set, our generative framework achieved a 20.0% top-1 exact match (1/5) and a 40.0% Hit@5 (2/5), maintaining a mean Jaccard similarity of 0.27 on the top-1 beam. This confirms that the model does not merely memorize literature conventions, but actively constructs viable synthetic pathways for genuinely unseen, highly complex materials

These real-world test cases illustrate both the robust generative capabilities of the model and the inherent limitations of evaluating synthesis planning via strict set-level matching. For the stoichiometric target $Li_2Mg_2(WO_4)_3$, the model successfully constructs the exact precursor combination by dynamically recombining fundamental structural motifs (e.g., WO_3, alkaline-earth/alkali carbonates, and simple oxides) into a coherent, target-specific recipe. Conversely, apparent exact-match "failures" frequently occur with doped or non-stoichiometric targets, where literature conventions for dopant fractions and vacancies remain highly ambiguous. In these instances, multiple oxide-route precursor combinations are chemically viable. Rather than failing, the model actively proposes scientifically sound alternatives (e.g., valid substitutions of rare-earth oxides), resulting in non-zero Jaccard similarities despite incurring severe

penalties under rigid Exact Match criteria. Furthermore, for complex targets such as $Ni_0 \cdot 5\, Zn_0 \cdot 5\, Fe_2O_4$, the exact ground-truth recipe is successfully generated within the top-5 candidate beams (Hit@5). This confirms that the model possesses the underlying capability to invent the correct out-of-corpus recipe, even when sequence-decoding heuristics (such as beam scoring and early EOS selection) rank it marginally below another chemically valid alternative.

Taken together, this prospective validation serves as a rigorous *in silico* proxy for bench-level experimental testing. The results confirm that our generative framework reliably constructs executable reaction pathways for *novel yet standard* solid-state targets specifically, unprecedented materials whose constituent stoichiometries and precursor families align with foundational inorganic paradigms (e.g., conventional oxide, carbonate, and binary routes). In these regimes, the model effectively leverages chemical priors to dynamically recombine building blocks into the exact experimental procedure. Crucially, when challenged with highly complex synthesis spaces, such as heavily doped, non-stoichiometric phases, or specialized techniques like flux growth and transport—the model does not fail randomly. Instead, it systematically generates thermodynamically plausible, mass-balanced alternative routes. The divergence from specific literature reports in these highly idiosyncratic spaces is not a failure of the model's chemical reasoning, but rather a reflection of the inherent multiplicity of viable physical synthesis pathways. Consequently, these generative outputs should be viewed not merely as textual approximations, but as robust, actionable synthetic hypotheses ready for physical laboratory execution.

Supplementary Table 2 Prospective external validation on out-of-corpus papers (in-vocabulary subset). Each row is a manually curated paper not present in train/test whose full ground-truth precursor set is in the $K = 150$ vocabulary. We report the target, manually extracted precursors, the top-1 prediction, set overlap (Jaccard), and whether the exact ground-truth set appears within the top-5 beams (Hit@5).

Source	Target	Manual precursors	Top-1 prediction	Jaccard	EM	Hit@5
[12]	$Gd_{0.995}Sm_{0.005}Al_{0.995}Cr_{0.005}O_3$	Gd_2O_3; Al_2O_3; Sm_2O_3; Cr_2O_3	Al_2O_3; Gd_2O_3; Yb_2O_3; Er_2O_3	0.33	No	No
[13]	$Ni_{0.5}Zn_{0.5}Fe_2O_4$	NiO; ZnO; Fe_2O_3	—	0.00	No	Yes
[14]	$La_{0.85}K_{0.10}MnO_3$	La_2O_3; K_2CO_3; $MnCO_3$	—	0.00	No	No
[15]	$Li_2Mg_2(WO_4)_3$	WO_3; Li_2CO_3; MgO	Li_2CO_3; MgO; WO_3	1.00	Yes	Yes
[12]	$Ba_{0.67}Ni_{0.33}MnFeO_3$	$BaCO_3$; NiO; MnO_2; Fe_2O_3	$BaCO_3$; Fe_2O_3; Mn_2O_3	0.40	No	No

References

[1] Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., Ceder, G.: Text-mined dataset of inorganic materials synthesis recipes. Scientific data **6**(1), 203 (2019)

[2] Lee, S., Cruse, K., Baibakova, V., Ceder, G., Jain, A.: Text-mined dataset of solid-state syntheses with impurity phases using large language model. Scientific Data (2025)

[3] Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., Ceder, G.: Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science **68**, 314–319 (2013)

[4] Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.: A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials **2**(1), 16028 (2016)

[5] Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., *et al.*: Matminer: An open source toolkit for materials data mining. Computational Materials Science **152**, 60–69 (2018)

[6] Gasteiger, J., Shuaibi, M., Sriram, A., Günnemann, S., Ulissi, Z., Zitnick, C.L., Das, A.: Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. arXiv preprint arXiv:2204.02782 (2022)

[7] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE transactions on big data **7**(3), 535–547 (2019)

[8] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[10] He, T., Huo, H., Bartel, C.J., Wang, Z., Cruse, K., Ceder, G.: Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. Science advances **9**(23), 8180 (2023)

[11] Vinyals, O., Bengio, S., Kudlur, M.: Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391 (2015)

[12] Tayari, F., Nassar, K.I., Algessair, S., Hjiri, M., Benamara, M.: Investigating fe-doped ba 0.67 ni 0.33 mn 1- x fe x o 3 (x= 0, 0.2) ceramics: Insights into electrical and dielectric behaviors. RSC advances **14**(18), 12561–12573 (2024)

[13] Elhamdi, I., Souissi, H., Kammoun, S., Dhahri, E., Pina, J., Costa, B., Brito, A., Fausto, R., López-Lago, E.: Structural and luminescent properties of a cr 3+/sm

3+ doped gdalo 3 orthorhombic perovskite for solid-state lighting applications. RSC advances **15**(3), 2066–2077 (2025)

[14] Vitayaya, O., Kurniawan, B., Nehan, P.Z.Z., Munazat, D.R., Sudiro, T., Imaduddin, A., Nugraha, H., Yudanto, S.D., Manawan, M.T.: Enhanced magnetoresistance properties of k-site deficient la 0.85 k 0.1 0.05 mno 3 manganites synthesized via sol–gel, wet-mixing, and solid-state reaction methods. RSC advances **14**(52), 38615–38633 (2024)

[15] Akermi, M., Mbarek, I., Hassani, R., Nasri, S., Oueslati, A.: Investigating li 2 mg 2 (wo 4) 3: structure, morphology, and electrical properties with ultra-low dielectric loss for optimizing laser host materials. RSC advances **15**(17), 13064–13075 (2025)