Language Models Enable Data-Augmented Synthesis Planning for Inorganic Materials

Thorben Prein[◇‡¶], **Elton Pan**[§], **Janik Jehkul**[◇], **Steffen Weinmann**[◇]
Elsa Olivetti[§], **Jennifer L. M. Rupp**[◇¶△*]

[◇] Technische Universität München, Arcisstraße 21, 80333 München, Germany
[‡] Munich Data Science Institute, Garching b. München, 85748, Germany
[¶] TUMint. Energy Research GmbH, Garching b. München, 85748, Germany
[§] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[△] Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany

Abstract

Inorganic synthesis planning has largely relied on heuristic strategies or machine-learning models trained on limited datasets, which restricts generality. We show that general-purpose language models, without task-specific fine-tuning, can recall synthesis conditions reported in the scientific literature. Off-the-shelf models, including *GPT-4.1*, *Gemini 2.0 Flash*, and *Llama 4 Maverick*, reach Top-1 precursor-prediction accuracy of up to 53.8% and Top-5 performance of 66.8% on a held-out set of 1,000 reactions. They also predict calcination and sintering temperatures with mean absolute errors below 126 °C, matching or surpassing specialized regression baselines. Ensembling these language models further improves predictive accuracy and cuts inference cost per prediction by up to 70%. Leveraging the broad, cross-domain knowledge of language models, we assess knowledge transfer by training a transformer, `SyntMTE`, on 28,548 LM-generated reaction recipes. Relative to a model trained on literature-reported data, a model trained solely on LM-generated data attains competitive performance (only 6% lower). Moreover, training on both LM-generated and literature-reported data yields up to a 4% improvement. In a case study on $Li_7La_3Zr_2O_{12}$ solid-state electrolytes, `SyntMTE` reproduces experimentally observed dopant-dependent sintering trends. Together, these results establish a hybrid workflow for scalable, data-efficient inorganic synthesis planning. This non-archival workshop paper summarizes work currently under review at ACS Applied Materials & Interfaces; portions of the text and figures are adapted from that manuscript.

1 Introduction

The discovery and design of advanced materials drive progress across energy conversion and storage, information technology, and medicine [1, 2, 3, 4, 5]. Recent advances in machine-learning–accelerated simulations have produced a rapid expansion of computationally predicted candidates, now numbering in the millions [6, 7]. Consequently, synthesizing these materials has emerged as the central bottleneck in the discovery pipeline [8, 9, 10, 11, 12]. While density functional theory provides valuable thermodynamic insight, accurately predicting kinetics, diffusion, and phase-transformation pathways remains challenging, leaving synthesis largely a trial-and-error endeavor [13, 14, 15, 16, 9]. To address the synthesis bottleneck, the community has increasingly turned to ML to mine protocols from the literature and to forecast feasible reaction routes for new compounds [17, 18,

*Corresponding author: `jrupp@tum.de`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) AI4Mat Workshop.

19]. Early contributions by Kononova et al. [16], E. Kim et al. [20], and Huo et al. [21] assembled large, curated corpora of synthesis procedures, laying the groundwork for ML-driven inorganic synthesis planning. Subsequent work has coalesced around two core subproblems: (i) precursor recommendation, selecting compatible reagent sets and (ii) synthesis-condition prediction, estimating operative parameters such as temperature and time. Chaining these steps yields end-to-end protocols for a specified target.

Precursor recommendation The majority of precursor recommendation techniques address solid-state synthesis. In one line of work, E. Kim et al.[22] leveraged an RNN with ELMo embeddings to extract over 50,000 actions and 116,000 precursor mentions, then trained a paired conditional VAE that jointly captures action sequences and precursor formulas to propose candidates for unseen targets. S. Kim[23] introduced element-wise retrosynthesis organized into 39 template classes. Retrieval-style models have also emerged: He et al.[24] used attention to compare planned routes against historical syntheses, and Noh et al.[25] incorporated an enthalpy-aware ranking step. Prein et al.[26] further enhanced generalization by embedding materials with a pretrained transformer and employing a pairwise ranker to score precursor sets for previously unseen compounds.

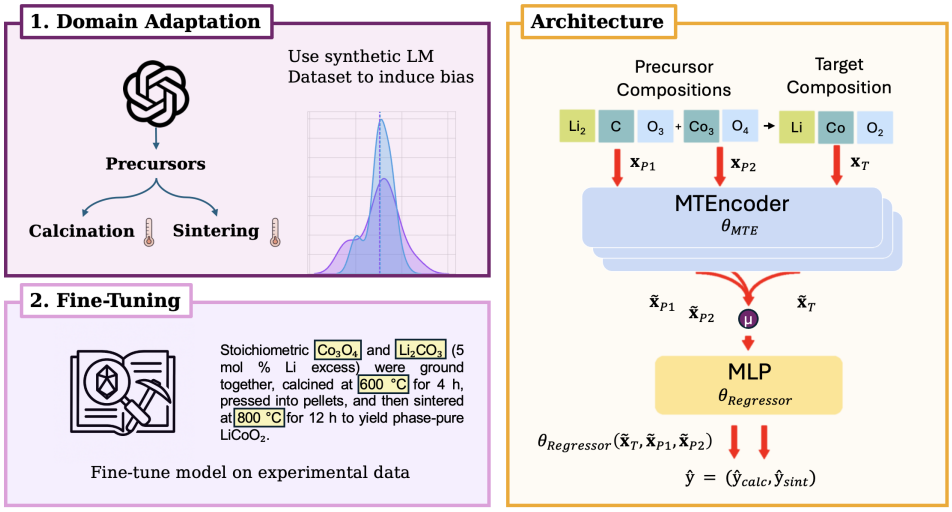

Figure 1: **Overview of our synthesis-condition model.** *Left:* We first adapt the MTEncoder on a large language-model–generated corpus to bias it toward solid-state reaction conditions, then fine-tune on experimental literature recipes. *Right:* Each precursor and the target composition are encoded with a shared MTEncoder (θ_{MTE}) into embeddings $\tilde{x}_{p,i}, \tilde{x}_T$, which are pooled and concatenated, then passed to an MLP head ($\theta_{\text{Regressor}}$) to predict calcination and sintering temperatures, $\hat{\mathbf{y}} = (\hat{y}_{\text{calc}}, \hat{y}_{\text{sint}})$.

Synthesis condition prediction Once precursors are selected, the next step is to predict isothermal-hold temperatures and dwell durations for calcination and sintering. Leveraging text-mined properties such as melting points and formation energies, Huo et al.[27] trained linear and tree-ensemble models and reported an MAE near 140°C. Prein et al.[28] introduced a Reaction Graph Network that couples MTEncoder representations with graph-attention. In parallel, Pan et al.[29, 30] formulated the task as conditional diffusion, using the target structure to generate plausible conditions and naturally reflecting the one-to-many structure–synthesis relationship.

Language models in materials synthesis Modern language models (LMs) are exposed during pretraining to vast amounts of unstructured chemical knowledge, including implicit heuristics, phase-diagram intuition, and procedural narratives captured across large text corpora. They have shown substantial utility across scientific domains [31, 32, 33, 34, 35]. In crystal-structure generation, generative LMs have achieved notable successes: CrystaLLM [36] and Crystal-Text-LLM [37] produce DFT-validated geometries, while FlowLLM [38] refines LLM-generated structures via flow matching [39]. For synthesis planning, GPT variants fine-tuned for synthesizability and precursor selection can approach the performance of specialized models [40]. Yet, systematic evaluation of state-of-the-art LMs on *both* precursor generation and processing-condition prediction for inorganic solid-state synthesis, the backbone of materials discovery has been lacking [12].

Establishing such a benchmark enables us to inform the community on model selection and clearer expectations of LM capabilities in this setting. We compare ensembles against single-model pipelines and analyze the distributions of predicted synthesis conditions. Beyond building a benchmark with current state-of-the-art models, we explore directions towards generating synthetic datasets via LMs, circumventing the data availability constraints in the domain. Finally, we also showcase our models applicability in a case study. To our knowledge, these aspects have not been comprehensively studied. This work addresses the following questions:

1. How well do state-of-the-art LMs perform on inorganic solid-state synthesis planning tasks?
2. Do LM ensembles outperform single models, and how do they affect the distribution and calibration of proposed recipes?
3. Can LM-generated synthesis recipes both enrich current literature-mined sparse databases and act as an informative prior for domain-specific models?

2 Methods

2.1 Benchmarking LMs in inorganic materials synthesis

To evaluate LMs for inorganic synthesis planning, we benchmark state-of-the-art systems on two tasks using a dataset derived from Kononova et al. [16] comprising $\sim 10{,}000$ unique precursor–target pairs. For precursor recommendation, we follow prior protocol [26] and construct a 1,000-example test set. Prompts are issued via OpenRouter without fixing the number of precursors, requiring each LM to infer how many reagents are appropriate for a given reaction. For synthesis-condition prediction, we curate a separate 1,000-entry subset by retaining records that report both calcination and sintering temperatures. Across both tasks, models receive 40 in-context examples drawn from a held-out validation split (Fig. 7). Precursor prediction is scored using exact-match accuracy [25]. We evaluate seven contemporary LMs to cover a diverse set of architectures; further model details appear in the Appendix.

2.2 LM ensembles

Based on validation performance, we assemble an ensemble of Gemini 2.0 Flash, Llama 4 Maverick, and DeepSeek Chatv3, and compare three rank-aggregation schemes: **Min-rank:** assign each candidate the best (lowest) rank it attains from any model, elevating items that at least one LM strongly supports. **Average-rank:** use the mean rank across models, balancing contributions and dampening the influence of outliers. **Max-rank:** assign each candidate the worst (highest) rank it receives, prioritizing only those consistently preferred by all models.

2.3 LM-enabled data augmentation

To generate a synthetic dataset, we queried the Materials Project [6] for entries with reported experimental syntheses, yielding 48,927 compounds. To promote broad chemical coverage, 10,000 target compositions were chosen using a maximum-entropy selection strategy. GPT–4.1 was prompted to generate precursor pathways and to filter out compositions unlikely to be accessible via solid-state routes. In line with prior practice, we retained the three highest-ranked suggestions per target (top-3 accuracy of 64.1%, Table 1). We then inferred processing parameters; after excluding incomplete generations, 29,473 records remained. Imposing minimum temperatures of $300^{\circ}\mathrm{C}$ for calcination and $500^{\circ}\mathrm{C}$ for sintering produced 28,548 candidate solid-state recipes. As summarized in Fig. 9a), the resulting collection spans a much wider compositional space than the literature-derived Kononova dataset [16].

2.4 SyntMTE architecture and training

To predict synthesis parameters, we introduce SyntMTE, which adapts the MTEncoder [41], pre-trained on the Alexandria DFT corpus across 12 properties [7], for actionable condition prediction from a reaction description. The target composition and each precursor are encoded with shared MTEncoder weights. The resulting embeddings are mean-pooled, concatenated, and fed to a two-layer MLP that performs multi-task regression of calcination and sintering temperatures. Training

Table 1: **Precursor recommendation performance.** Top-k exact-match accuracies for individual language models and ensemble strategies on retrosynthesis precursor prediction. GPT-4.1 achieves the highest Top-1 accuracy, while min-rank ensembles boost performance at higher Top-k thresholds. Notably, the ensemble of Llama 4 Maverick, DeepSeek Chat v3, and Gemini 2.0 Flash surpasses GPT-4.1 for relevant Top-5 and Top-10 settings with a 70% reduction in inference cost.

Model	Top-1 ↑	Top-3 ↑	Top-5 ↑	Top-10 ↑
Ensemble Min-Rank	52.3	**65.8**	**70.7**	**74.3**
Ensemble Min-Rank	51.8	63.1	67.4	71.9
OpenAI GPT-4.1	**53.8**	**64.1**	66.1	68.7
Grok 3 Mini Beta	52.2	63.2	**66.8**	**69.5**
Llama 4 Maverick	53.1	61.1	64.2	69.3
DeepSeek Chat v3	53.5	60.7	63.7	66.2
Mistral Small 3.1	52.0	59.7	61.7	63.9
Gemini 2.0 Flash-001	51.4	59.2	62.0	66.2
Qwen 2.5 VL	50.7	55.5	58.0	59.3

follows a chronological split on the Kononova corpus [16] (train $\leq$2014, validation 2015–2016, test $>$2016) to emulate forward-in-time deployment, and uses an LM-generated synthetic corpus for broad coverage before final alignment on literature yielding realistic, lab-ready temperature predictions.

3 Results and Discussion

3.1 Precursor recommendation

We assess the language models on the precursor prediction task and report Top-k exact-match accuracy. This metric is a conservative estimate of performance because it requires the model to reproduce exactly the precursor set documented in the literature, even though other valid, undocumented synthesis routes may exist [29]. Since practical precursor selection involves testing several candidate routes, the Top-5 and Top-10 results are especially useful, they show whether a correct set appears among the model's five or ten highest-ranked suggestions. As summarized in Table 1, all models perform competitively and fall within a tight range. Qwen 2.5 VL is the only clear underperformer despite its substantial parameter count. OpenAI GPT-4.1 ranks first for Top-1 at 53.8% and maintains strong results as k increases, followed by Grok 3 mini, Llama 4 Maverick, and DeepSeek Chat v3. We compare LM outcomes with literature baselines, noting that the comparison is imperfect: baseline systems were trained on smaller datasets, and LMs could benefit from pretraining exposure to test-set synthesis protocols. The best reported baseline achieves Top-5 and Top-10 accuracies of 73% and 78% [26], whereas individual LMs approach these figures with up to 66% and 69%. Overall, this highlights that state-of-the-art LMs, without chemistry-specific objectives, can retrieve high-quality chemical knowledge via in-context learning alone.

LM ensemble We find that the Min-rank and Average-rank schemes markedly boost performance at Top-3, Top-5, and Top-10 (Fig. 5), with only a slight decrease in Top-1 accuracy relative to the best single model. The ensemble's strong recall is driven by its diversity. In the information-retrieval literature, rank-fusion methods consistently raise recall by leveraging complementary strengths of heterogeneous rankers across topics and queries, and increased diversity is linked to improved recall in ranking tasks [42, 43]. Accordingly, min-rank and average-rank aggregation are well suited to our LM ensembles: they promote any candidate highly ranked by at least one model into the final Top-k, thereby maximizing recall.

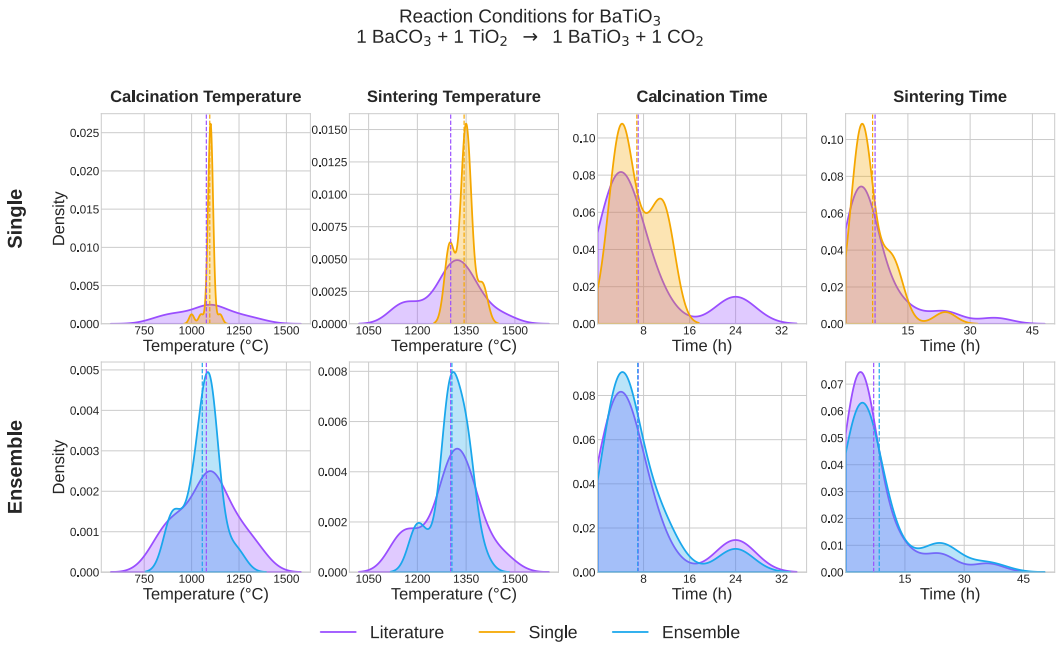

Figure 2: **Synthesis condition distributions of literature-reported and LM generated solid-state synthesis recipes for $BaTiO_3$.** Literature distribution KDEs are shaded purple. Dotted lines refer to the mean value. 'Single' refers to LM distributions acquired by drawing 24 samples from Gemini-2.0-Flash (orange). 'Ensemble' refers to LM distributions acquired by drawing 8 predictions each are sampled from Llama Maverick, DeepSeek Chat v3, and Gemini 2.0 Flash (blue). The individual LMs yield narrower distributions that fail to capture the underlying literature distribution, whereas the ensemble more accurately reproduces the literature's secondary modes.

3.2 Synthesis condition regression

For synthesis-condition regression, we assess LM performance on predicting parameters for a standard solid-state route comprising two heating stages. First, precursor powders are mixed and homogenized to achieve uniformity. Then, during calcination, the blended precursors are heated so that thermal decomposition and diffusion initiate formation of the target phase. Finally, in the sintering stage, elevated temperatures enhance grain-boundary and volume diffusion, leading to neck formation and growth that consolidate and densify the powder into a cohesive bulk body [44, 45]. To mirror experimental practice, we prompt the LMs to output calcination and sintering temperatures and evaluate the predictions against the curated 1,000-entry subset from Kononova [16]. We exclude dwell times because they are highly operator-dependent and therefore difficult to model via regression [9, 27]. As a consequence, prior point-regression approaches often captured researcher-specific noise rather than underlying thermodynamics, yielding low R^2 and weak predictive performance [9, 27]. Moreover, for regression-style problems, it is important to remember that LMs are fundamentally next-token predictors optimized with a classification objective, and do not possess built-in, high-precision numerical reasoning. Consequently, they are generally a poor fit for pure regression [46]. That said, several studies have demonstrated that LMs can tackle challenging numerical tasks with notable success [47]. In practical synthesis reporting, temperatures are almost always given as integer values (e.g., 800 °C), which further facilitates LM applicability. Table 2 summarizes our results on the synthesis-condition regression task. For calcination temperature prediction, OpenAI GPT-4.1 attains the best performance, followed by Gemini 2.0 Flash and DeepSeek Chat v3. For sintering temperature, Gemini 2.0 Flash leads, with Llama 4 Maverick and OpenAI GPT-4.1 next in line. Grok 3 Mini Beta, previously second on precursor prediction, places last on both regression evaluations. Overall, model rankings are broadly correlated across the two tasks.

Single-model methods reach MAEs of 101 °C for sintering temperature prediction, compared with 127 °C for calcination temperature regression. This gap is notable given that sintering temperatures are generally higher in absolute magnitude. We find, that the calcination temperature distribution

Table 2: **Synthesis condition prediction performance.** Regression performance for calcination and sintering temperature prediction. **Bold** indicates the best and underlined the second-best value in each metric. Ensemble 1 comprises Gemini 2.0 Flash, Llama 4 Maverick, and DeepSeek Chat v3, Ensemble 2 features OpenAI GPT-4.1, Gemini 2.0 Flash, and DeepSeek Chat v3.

Model	Sintering temperature			Calcination temperature		
	MAE (↓)	R^2 (↑)	RMSE (↓)	MAE (↓)	R^2 (↑)	RMSE (↓)
Ensemble Avg	**96.31**	**0.667**	**134.48**	125.72	0.410	168.86
Ensemble Avg	96.89	0.6627	135.42	**123.00**	**0.424**	**166.93**
Gemini 2.0 Flash-001	**100.66**	**0.628**	**142.22**	127.04	0.356	176.53
Llama 4 Maverick	102.76	0.612	145.23	135.85	0.323	180.90
OpenAI GPT-4.1	105.21	0.586	150.01	**125.92**	**0.371**	**174.45**
DeepSeek Chat v3	106.40	0.610	145.73	132.48	0.309	182.78
Mistral Small 3.1	113.93	0.550	156.36	137.05	0.291	185.20
Qwen 2.5 VL	131.93	0.443	174.06	142.68	0.232	192.72
Grok-3 Mini Beta	131.00	0.433	175.56	152.09	0.123	205.97

has higher variance, about 16% higher normalized spread than sintering. Additionally, calcination temperature appears more sensitive to factors not captured in the dataset (e.g., precursor particle size), which are known to influence calcination conditions. For instance, Pavlović *et al.* show that increasing the ball-milling time for $BaTiO_3$ by one hour can lower the required calcination temperature by more than 100 °C [48]. Overall reinforcing calcination temperature prediction as the more challenging task.

LM ensemble Echoing the precursor recommendation setup, we construct simple three-model ensembles by averaging LM predictions. For sintering-temperature regression, combining Llama 4 Maverick, Gemini 2.0 Flash, and DeepSeek Chat v3 delivers a 4% absolute gain in R^2 over the strongest single model. For calcination-temperature regression, a second ensemble that replaces Llama with GPT-4.1 improves R^2 by 5%, reaching 42.4%. Consistent with the precursor task, these ensemble configurations both boost performance and may reduce inference cost by roughly 70%.

We explain why LM ensembles surpass single models. In materials synthesis, the mapping from process recipes to a target compound is inherently one-to-many[29]. A given composition, e.g., $BaTiO_3$, can be obtained via multiple annealing protocols that differ in calcination and sintering settings, especially temperature and dwell time. We generate LM-based distributions over synthesis conditions and compare them with literature-reported protocols for 24 pristine $BaTiO_3$ samples. As illustrated in Fig. 2, individual LMs produce narrow distributions with a single dominant mode sharply centered near the mean (orange, top row). In contrast, an LM ensemble exhibits much better overlap with the ground truth (purple). For calcination temperature, for instance, it recovers a secondary mode below the mean while also matching the primary mode (blue, bottom row). Likewise, for sintering temperature, the ensemble distribution captures the mean and additional structure in the target distribution, including a regime around 1200 °C. Most notably, for synthesis time, single LMs regress toward the mean with limited spread, whereas the ensemble more faithfully reflects the behavior at longer durations. This shows the efficient modelling of one-to-many synthesis relationships via LM ensembles.

3.3 LM ensembles expand the performance-cost pareto front

To compare overall LM performance, we normalize each model's score to the best-performing model and then average these normalized values. We estimate inference cost using input/output token pricing (Fig. 3). GPT-4.1 and Gemini deliver the highest mean normalized performance; however, Gemini's much lower price point makes it particularly appealing for materials-informatics use. Notably, an ensemble of lower-cost models, Llama 4 Maverick, DeepSeek Chat v3, and Gemini 2.0 Flash, surpasses any single model while cutting cost by roughly 70% relative to GPT-4.1. In addition,

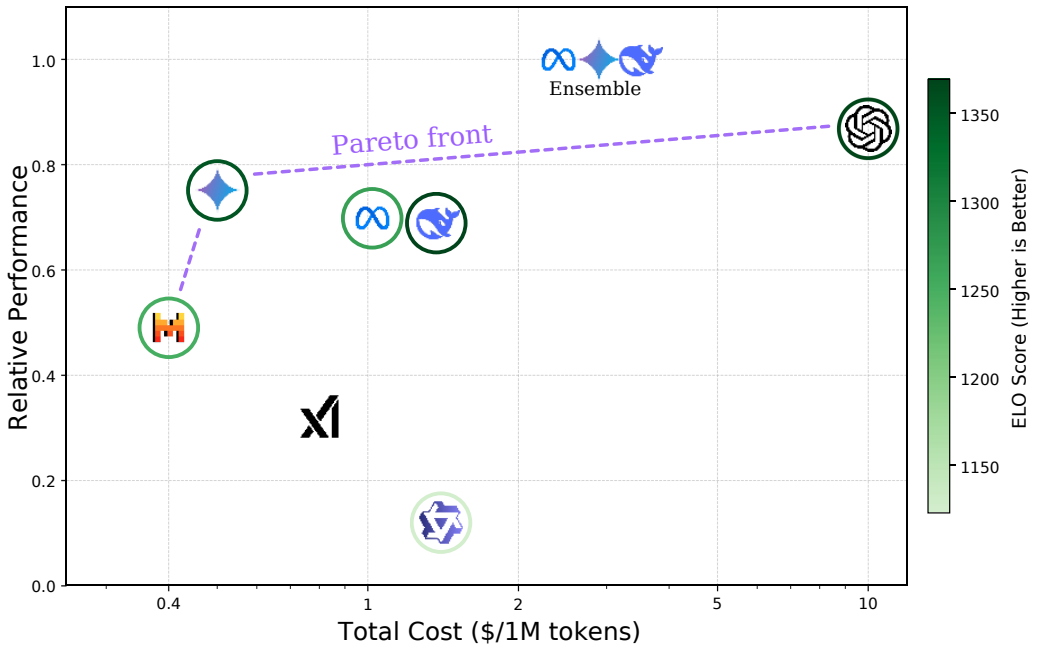

Figure 3: **Comparison of model performance vs. cost.** We compute each model's relative performance on precursor prediction, calcination and sintering temperature estimation tasks, and plot the average performance relative to cost. GPT-4.1 delivers highest individual performance and comes at the highest cost. An ensemble of Llama 4 Maverick, DeepSeek Chat v3 and Gemini 2.0 Flash surpasses any single model in performance while reducing cost by 70% relative to GPT-4.1. The ELO rating score is represented by the color of each circle and serves as a quantitative indicator of model performance [49].

our analyses show that ensembles produce output distributions that more closely track the scientific literature, reinforcing their combined advantages in cost and performance.

3.4 Synthetic data augmentation improves model performance

We study how datasets expanded with LM outputs affect leading methods for predicting synthesis conditions. Our setup follows NLP practice, where limited, specialized corpora are bolstered with LM-generated text (e.g., Xu et al. [50]) or with teacher–student style pseudo-labels [51, 52, 53]. The core idea is to leverage an LM's prior over synthesis parameters to initialize smaller expert models: we first fit to LM-derived estimates to capture broad regularities, and then continue training on experimental observations. Building on this workflow, we introduce `SyntMTE`, a composition-focused model derived from MTEncoder, a transformer representation for inorganic materials pretrained on the large-scale Alexandria DFT collection [41, 7]. Pretraining on extensive DFT data improves representations and yields better results across many materials tasks. As in NLP and vision, the pretraining objective can be only loosely related to the downstream goal; broad, physics-based supervision still shapes the internal chemical-space representation in ways that transfer effectively. We therefore exploit the scale and diversity of public DFT datasets, covering millions of computed properties, to pretrain MTEncoder and subsequently fine-tune it for synthesis prediction, which consistently outperforms training from scratch.

Our approach extends over previous work [9] by embedding not only the reaction products but also the full set of precursor materials. This makes the task more challenging: a single target compound can be paired with multiple precursor combinations, each implying different processing temperatures. At the same time, it better reflects practical solid-state synthesis. After mapping every material participating in a reaction to an embedding, we form a reaction vector via mean pooling and predict the processing parameters with a multi-task regression head (Figure 1, right). For evaluation, we use the Kononova dataset [16] with a time-aware split, training on entries up to 2015, validating on 2015–2016, and

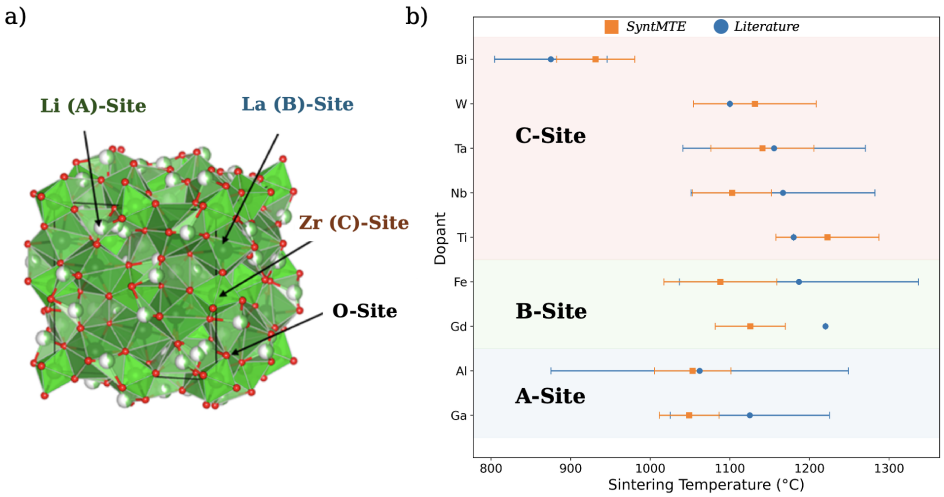

Figure 4: **LLZO-case study.** a) Probable doping sites in the cubic LLZO unit cell. Reproduced from ref mahbub2020text. Available under a CC BY-NC-ND 4.0 license (`https://creativecommons.org/licenses/by-nc-nd/4.0/`). Copyright 2020 Mahbub et al. b) True (blue circles) vs. predicted (orange squares) sintering temperatures with mean and standard deviation across different case reports per cation of the garnet electrolyte. Dopants grouped by crystallographic substitution site.

testing on later publications. We compare against three baselines: (i) a composition-only feedforward network, (ii) a CrabNet-style transformer [55], and (iii) an XGBoost model operating on mean-pooled reaction features. We study three training schemes: two-stage fine-tuning (synthetic data followed by literature data), direct fine-tuning on literature only, and training solely on synthetic data.

The results are reported in Tab. 3. Training on synthetic recipes alone delivers strong performance for all models despite having no exposure to literature-mined samples and no overlap with the literature test set. The coefficient of determination (R^2) remains high across models, with the exception of XGBoost. This drop is attributable to XGBoost's single-task formulation: it fits calcination and sintering separately and thus misses the inductive bias of joint (multi-task) learning, which especially degrades calcination temperature predictions. We then contrast the two fine-tuning approaches. For every model except XGBoost, incorporating synthetic data improves accuracy, as shown by the relative MAE reductions in Table 3. The largest gains appear for `SyntMTE`, which lowers MAE by over 4% on both targets. CrabNet also improves (a 1% MAE reduction), even narrowly surpassing the `SyntMTE` variant trained solely on experimental data. Overall, representation-learning methods, `SyntMTE` and CrabNet, benefit most from augmentation. This is illustrated in Figure 6, which compares parity plots for two `SyntMTE` models: one trained only on literature data and one trained with the augmented set. In sum, comparing literature-only training with training that includes LM-generated synthetic data underscores the utility of synthetic corpora for synthesis modeling. Models trained exclusively on large synthetic sets still achieve competitive accuracy, reducing the need for labor-intensive manual literature curation. Furthermore, `SyntMTE`'s DFT-based pretraining is particularly advantageous when literature data are scarce. Finally, when set against the top ensembles in the LM benchmark (Table 2), our expert models outperform on all regression targets. Because benchmark scores may be inflated by data leakage, we advise caution in comparing them to our year-split evaluations; even so, the discrepancy highlights the rapid advancement of LM capabilities.

3.5 Case study

Virtual screening of compound-specific sintering temperatures and durations provides a quantitative proxy for manufacturing cost [56]. As a case study, we evaluate garnet-type LLZO ($Li_7La_3Zr{*}2O{*}12$), a leading oxide solid-state electrolyte valued for high ionic conductivity but constrained by costly, high-temperature densification [57, 58, 59, 60, 61]. We curate 40 literature synthesis routes of doped LLZO and test `SyntMTE` in a strict extrapolation setting by withholding all LLZO-like compositions from training. Because multiple viable sintering protocols exist per dopant family, we assess

qualitative ordering rather than exact temperatures. Across C-site dopants, the model reproduces known trends: Ta^{5+} substitution (charge-balanced by V_{Li}) yields moderate densification aid with experimental pellets sintered at 1100–1150°C, `SyntMTE` recovers a matching window with mean near reported values and spread 1060–1200°C [62, 63, 64, 65]. Bi^{3+} substitution predicts a narrow, lower-temperature window (model: 880–980°C) and captures the sharp decline to 850°C observed experimentally [66, 67]. For A/B-site dopants, Al exhibits a broad window (small radius, mixed-site occupancy), whereas Ga yields a narrower profile, predictions for Fe align well with reports, while Gd is overly optimistic [68, 69]. Overall, despite no prior LLZO training, `SyntMTE` recovers key sintering-temperature trends, indicating that synthesis-planning models can guide dopant selection and reduce processing temperatures and thus costs via virtual screening.

4 Conclusions

Machine-learning approaches to materials synthesis are constrained by limited data, we show that modern language models can mitigate this. Across seven models evaluated on precursor recommendation and processing-parameter regression, top-1 exact-match exceeds 50% (66% top-5). Ensembles better capture multimodal synthesis windows and can cut inference cost by up to 70%. We distill literature knowledge into 28,548 synthetic solid-state recipes and use it for two-stage training of `SyntMTE` (synthetic pretraining, then literature fine-tuning). This yields state-of-the-art results, surpassing CrabNet and lowers MAE for sintering and calcination by about 6°C. In a LLZO case study, the model recovers broad sintering windows and the temperature drop from Bi substitution, suggesting practical guidance for lower-temperature routes. Overall, LM-based augmentation offers scalable, low-cost auxiliary data on commonly reported synthesis variables and can pair with Bayesian optimization and autonomous experimentation to accelerate discovery and scale-up of advanced materials.

References

[1] Huanping Zhou et al. “Interface engineering of highly efficient perovskite solar cells”. In: *Science* 345.6196 (2014), pp. 542–546.

[2] Maarten de Jong et al. “The high-throughput highway to computational materials design”. In: *Scientific Data* 2 (2013), p. 150009.

[3] George M Whitesides. “The’right’size in nanobiotechnology”. In: *Nature biotechnology* 21.10 (2003), pp. 1161–1165.

[4] John B Goodenough and Youngsik Kim. “Challenges for rechargeable batteries”. In: *Journal of Power Sources* 196.16 (2011), pp. 6688–6694.

[5] Yi Cui et al. “High performance silicon nanowire field effect transistors”. In: *Nano letters* 3.2 (2003), pp. 149–152.

[6] Anubhav Jain et al. “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL materials* 1.1 (2013), p. 011002.

[7] Jonathan Schmidt et al. “Improving machine-learning models in materials science through large datasets”. In: *Materials Today Physics* 48 (2024), p. 101560.

[8] Matthew J. McDermott et al. “Assessing thermodynamic selectivity of solid-state reactions for the predictive synthesis of inorganic materials”. In: *ACS Central Science* 9.10 (2023), pp. 1957–1975.

[9] Christopher Karpovich et al. “Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction”. In: *Chemistry of Materials* 35.3 (2023), pp. 1062–1079.

[10] Jehad Abed et al. “Open Catalyst Experiments 2024 (OCx24): Bridging Experiments and Computational Models”. In: *arXiv preprint arXiv:2411.11783* (2024).

[11] Shreshth A Malik, Rhys EA Goodall, and Alpha A Lee. “Predicting the outcomes of material syntheses with deep learning”. In: *Chemistry of Materials* 33.2 (2021), pp. 616–624.

[12] Nathan J Szymanski et al. “An autonomous laboratory for the accelerated synthesis of novel materials”. In: *Nature* 624.7990 (2023), pp. 86–91.

[13] Filip T Szczypiński, Steven Bennett, and Kim E Jelfs. “Can we predict materials that can be synthesised?” In: *Chemical Science* 12.3 (2021), pp. 830–840.

[14] Rachel Woods-Robinson et al. “Role of disorder in the synthesis of metastable zinc zirconium nitrides”. In: *Physical Review Materials* 6.4 (2022), p. 043804.

[15] Jiadong Chen et al. “Navigating phase diagram complexity to guide robotic inorganic materials synthesis”. In: *Nature Synthesis* 3.5 (2024), pp. 606–614.

[16] Olga Kononova et al. “Text-mined dataset of inorganic materials synthesis recipes”. In: *Scientific data* 6.1 (2019), pp. 1–11.

[17] Elsa A Olivetti et al. “Data-driven materials research enabled by natural language processing and information extraction”. In: *Applied Physics Reviews* 7.4 (2020).

[18] Olga Kononova et al. “Opportunities and challenges of text mining in materials research”. In: *Iscience* 24.3 (2021).

[19] Wenhao Sun and Nicholas David. “A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes”. In: *Faraday Discussions* 256 (2025), pp. 614–638.

[20] Edward Kim et al. “Materials synthesis insights from scientific literature via text extraction and machine learning”. In: *Chemistry of Materials* 29.21 (2017), pp. 9436–9444.

[21] Haoyan Huo et al. “Semi-supervised machine-learning classification of materials synthesis procedures”. In: *Npj Computational Materials* 5.1 (2019), p. 62.

[22] Edward Kim et al. “Inorganic materials synthesis planning with literature-trained neural networks”. In: *Journal of chemical information and modeling* 60.3 (2020), pp. 1194–1201.

[23] Seongmin Kim et al. “Element-wise formulation of inorganic retrosynthesis”. In: *AI for Accelerated Materials Design NeurIPS 2022 Workshop*. 2022.

[24] Tanjin He et al. “Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature”. In: *Science advances* 9.23 (2023), eadg8180.

[25] Heewoong Noh et al. “Retrieval-Retro: Retrieval-based Inorganic Retrosynthesis with Expert Knowledge”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 25375–25400.

[26] Thorben Prein et al. “Retro-Rank-In: A Ranking-Based Approach for Inorganic Materials Synthesis Planning”. In: *arXiv preprint arXiv:2502.04289* (2025).

[27] Haoyan Huo et al. “Machine-learning rationalization and prediction of solid-state synthesis conditions”. In: *Chemistry of Materials* 34.16 (2022), pp. 7323–7336.

[28] Thorben Prein et al. “Reaction Graph Networks for Inorganic Synthesis Condition Prediction of Solid State Materials”. In: *AI for Accelerated Materials Design—NeurIPS 2024 Workshop*. 2024.

[29] Elton Pan et al. “A Chemically-Guided Generative Diffusion Model for Materials Synthesis Planning”. In: *AI for Accelerated Materials Design—NeurIPS 2024 Workshop*. 2024.

[30] Elton Pan et al. “ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters”. In: *ACS Central Science* 10.3 (2024), pp. 729–743.

[31] Adrian Mirza et al. “A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists”. In: *Nature Chemistry* (2025), pp. 1–8.

[32] Juraj Gottweis et al. “Towards an AI co-scientist”. In: *arXiv preprint arXiv:2502.18864* (2025).

[33] José R Penadés et al. “AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution”. In: *bioRxiv* (2025), pp. 2025–02.

[34] Kai-Hendrik Cohrs et al. “Large language models for causal hypothesis generation in science”. In: *Machine Learning: Science and Technology* 6.1 (2025), p. 013001.

[35] Ross Taylor et al. “Galactica: A large language model for science”. In: *arXiv preprint arXiv:2211.09085* (2022).

[36] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. “Crystal structure generation with autoregressive large language modeling”. In: *Nature Communications* 15.1 (2024), pp. 1–16.

[37] Andre Niyongabo Rubungo et al. “LLM4Mat-bench: benchmarking large language models for materials property prediction”. In: *arXiv preprint arXiv:2411.00177* (2024).

[38] Anuroop Sriram et al. “FlowLLM: Flow matching for material generation with large language models as base distributions”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 46025–46046.

[39] Benjamin Kurt Miller et al. “Flowmm: Generating materials with riemannian flow matching”. In: *Forty-first International Conference on Machine Learning*. 2024.

[40] Seongmin Kim, Yousung Jung, and Joshua Schrier. “Large language models for inorganic synthesis predictions”. In: *Journal of the American Chemical Society* 146.29 (2024), pp. 19654–19659.

[41] Thorben Prein et al. “MTENCODER: A Multi-task Pretrained Transformer Encoder for Materials Representation Learning”. In: *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*. 2023.

[42] Yanzhao Wu et al. “Promoting high diversity ensemble learning with ensemblebench”. In: *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE. 2020, pp. 208–217.

[43] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. “Reciprocal rank fusion outperforms condorcet and individual rank learning methods”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 758–759.

[44] W David Kingery, Harvey Kent Bowen, and Donald R Uhlmann. *Introduction to ceramics*. John wiley & sons, 1976.

[45] James Stalford Reed. “Principles of ceramics processing”. In: (1995).

[46] Jonas Zausinger et al. “Regress, Don’t Guess–A Regression-like Loss on Number Tokens for Language Models”. In: *arXiv preprint arXiv:2411.02083* (2024).

[47] Karl Cobbe et al. “Training Verifiers to Solve Math Word Problems”. In: *arXiv preprint arXiv:2110.14168* (2021).

[48] Vera P Pavlović et al. “Synthesis of BaTiO3 from a mechanically activated BaCO3-TiO2 system”. In: *Science of Sintering* 40.1 (2008), pp. 21–26.

[49] Wei-Lin Chiang et al. “Chatbot arena: An open platform for evaluating llms by human preference”. In: *Forty-first International Conference on Machine Learning*. 2024.

[50] Benfeng Xu et al. “S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 8186–8207.

[51] Taha ValizadehAslani et al. “Two-stage fine-tuning with ChatGPT data augmentation for learning class-imbalanced data”. In: *Neurocomputing* 592 (2024), p. 127801.

[52] Tom Pieper et al. “Enhancing Small Language Models via ChatGPT and Dataset Augmentation”. In: *International Conference on Applications of Natural Language to Information Systems*. Springer. 2024, pp. 269–279.

[53] Minh Tran et al. “A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic data”. In: *arXiv preprint arXiv:2501.12501* (2025).

[54] Rubayyat Mahbub et al. “Text mining for processing conditions of solid-state battery electrolytes”. In: *Electrochemistry Communications* 121 (2020), p. 106860.

[55] Anthony Yu-Tung Wang et al. “Compositionally restricted attention-based network for materials property predictions”. In: *Npj Computational Materials* 7.1 (2021), p. 77.

[56] Christian Bauer et al. “Charging sustainable batteries”. In: *Nature Sustainability* 5.3 (2022), pp. 176–178.

[57] Moran Balaish et al. “Processing thin but robust electrolytes for solid-state batteries”. In: *Nature Energy* 6.3 (2021), pp. 227–239.

[58] Moran Balaish et al. “Emerging Processing Guidelines for Solid Electrolytes in the Era of Oxide-Based Solid-State Batteries”. In: (2025). Manuscript under review.

[59] Kun Joong Kim et al. “Solid-state Li–metal batteries: challenges and horizons of oxide and sulfide solid electrolytes and their interfaces”. In: *Advanced Energy Materials* 11.1 (2021), p. 2002689.

[60] Sara Catherine Sand, Jennifer LM Rupp, and Bilge Yildiz. “A critical review on Li-ion transport, chemistry and structure of ceramic–polymer composite electrolytes for solid state batteries”. In: *Chemical Society Reviews* (2024).

[61] Steffen Weinmann et al. “Sustainable functional ceramics”. In: *Nature Nanotechnology* (2025). Manuscript under review.

[62] Siqi Li et al. “Understanding the defect chemistry and Li+ transportation of Ta-doped Li7La3Zr2Ta0. 5O12-δ by active ML learning Raman spectroscopy image”. In: *Proceedings of 24th International Conference on Solid State Ionics (SSI24)*. 2024.

[63] Anatolii V Morozov et al. “Thermodynamics as a Driving Factor of LiCoO2 Grain Growth on Nanocrystalline Ta-LLZO Thin Films for All-Solid-State Batteries”. In: *ACS Applied Materials & Interfaces* 14.35 (2022), pp. 39907–39916.

[64] Fudong Han et al. “Electrochemical stability of Li10GeP2S12 and Li7La3Zr2O12 solid electrolytes”. In: *Advanced Energy Materials* 6.8 (2016), p. 1501590.

[65] Sewon Kim et al. “The role of interlayer chemistry in Li-metal growth through a garnet-type solid electrolyte”. In: *Advanced Energy Materials* 10.12 (2020), p. 1903993.

[66] Derek K Schwanz et al. “Bi aliovalent substitution in Li7La3Zr2O12 garnets: Structural and ionic conductivity effects”. In: *Aip Advances* 10.3 (2020).

[67] Reinhard Wagner et al. “Synthesis, crystal structure, and stability of cubic Li7–x La3Zr2–x Bi x O12”. In: *Inorganic chemistry* 55.23 (2016), pp. 12211–12219.

[68] Janez Košir et al. “Supervalent doping and its effect on the thermal, structural and electrochemical properties of Li 7 La 3 Zr 2 O 12 solid electrolytes”. In: *Materials Advances* 5.12 (2024), pp. 5260–5274.

[69] Alfred Junio Samson et al. “A bird’s-eye view of Li-stuffed garnet-type Li 7 La 3 Zr 2 O 12 ceramic electrolytes for advanced all-solid-state Li batteries”. In: *Energy & Environmental Science* 12.10 (2019), pp. 2957–2975.

[70] Shuai Bai et al. “Qwen 2.5-VL Technical Report”. In: *arXiv preprint arXiv:2502.13923* (2025). cs.CV. eprint: 2502.13923. URL: https://arxiv.org/abs/2502.13923.

[71] LMArena. “LMArena Leaderboard Overview”. In: *Online Resource* (2025). Accessed 24 May 2025. URL: https://beta.lmarena.ai/leaderboard.

[72] Mistral AI. “Mistral Small 3.1”. In: (2025). Accessed 2025-05-18. https://mistral.ai/news/mistral-small-3-1.

[73] DeepSeek-AI et al. “DeepSeek-V3 Technical Report”. In: *arXiv preprint* arXiv:2412.19437 (2025). arXiv: 2412.19437 [cs.CL]. URL: https://arxiv.org/abs/2412.19437.

[74] DeepSeek. “DeepSeek-V3-0324 Release”. In: (2025). Accessed: 2025-05-24.

[75] OpenAI. “GPT-4: A Sneak Peek at OpenAI’s Next Generation Model”. In: (2024). Accessed 2025-05-18. https://openai.com/index/gpt-4-1/.

[76] Meta AI. “Llama-4: Multimodal Intelligence”. In: (2024). Accessed 2025-05-18. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[77] xAI. “Grok-3”. In: (2024). Accessed 2025-05-18. https://x.ai/news/grok-3.

[78] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[79] Atsushi Togo, Kohei Shinohara, and Isao Tanaka. “Spglib: a software library for crystal symmetry search”. In: *Science and Technology of Advanced Materials: Methods* 4.1 (2024), p. 2384822.

[80] B Morosin. “The crystal structure of copper (II) tetraammine nitrate”. In: *Structural Science* 32.4 (1976), pp. 1237–1240.

[81] Rhys EA Goodall and Alpha A Lee. “Predicting materials properties without crystal structure: Deep representation learning from stoichiometry”. In: *Nature communications* 11.1 (2020), pp. 1–9.

[82] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.