

# Large Language Models for Superconductor Discovery

Suman Itani<sup>1,†</sup>, Yibo Zhang<sup>1,†</sup>, Ranjit Itani<sup>2</sup>, and Jiadong Zang<sup>1</sup>

<sup>1</sup>*Department of Physics and Astronomy, University of New Hampshire, 9 Library Way, Durham, NH 03824, USA*

<sup>2</sup>*Department of Electrical and Computer Engineering,  
University of New Hampshire, 33 Academic Way, Durham, NH 03824, USA\**

(Dated: December 12, 2025)

Large language models (LLMs) offer new opportunities for automated data extraction and property prediction across materials science, yet their use in superconductivity research remains limited. Here we construct a large experimental database of 78,203 records, covering 19,058 unique compositions, extracted from scientific literature using LLM driven workflow. Each entry includes chemical composition, critical temperature, measurement pressure, structural descriptors, and critical fields. We fine-tune several open-source LLMs for three tasks: (i) classifying superconductors vs. non-superconductors, (ii) predicting the superconducting transition temperature directly from composition or structure-informed inputs, and (iii) inverse design of candidate compositions conditioned on target  $T_c$ . The fine-tuned LLMs achieve performance comparable to traditional feature-based models—and in some cases exceed them—while substantially outperforming their base versions and capturing meaningful chemical and structural trends. The inverse-design model generates chemically plausible compositions, including 28% novel candidates not seen in training. Finally, applying the trained predictors to GNOME database identifies unreported materials with predicted  $T_c > 10K$ . Although unverified, these candidates illustrate how integrating LLM-driven workflow can enable scalable hypothesis generation for superconductivity discovery.

## I. INTRODUCTION

Superconductivity is a quantum phenomenon in which a material exhibits zero electrical resistance and expels magnetic fields below a critical temperature ( $T_c$ ). The relentless pursuit of superconductors with higher transition temperatures is driven by their transformative potential in energy-efficient power transmission, high-field magnets, magnetic levitation, and quantum technologies [1, 2]. Realizing these applications requires materials that maintain superconductivity at high  $T_c$  under practical and affordable cooling conditions. To date, only a few material families—primarily cuprates [3] and iron-based superconductors at ambient pressure [4], and hydrogen-rich hydrides at extreme pressures exceeding 150 GPa [5]—have demonstrated  $T_c$  values above 100 K. Achieving room-temperature superconductivity at ambient pressure remains one of the most compelling and unresolved challenges in condensed matter physics and materials science.

Identifying new superconductors remains challenging due to the absence of a universal predictive theory. The Bardeen–Cooper–Schrieffer (BCS) theory [6, 7] successfully explains phonon-mediated superconductivity, but many high- $T_c$  materials exhibit strong electronic correlations or unconventional pairing mechanisms that lie beyond the scope of BCS theory. Solving the full many-body Schrödinger equation for realistic materials is computationally infeasible, while density functional theory often underestimates key interactions in strongly correlated systems. As a result, superconductor discovery has historically relied on empirical rules and trial-and-

error synthesis, with only a small fraction of chemically accessible compounds exhibiting superconductivity [8]. Given the vast compositional and structural design space, purely heuristic approaches alone are no longer sufficient. Data-driven techniques offer a promising complementary pathway by uncovering patterns embedded in known superconductors that may accelerate the identification of new candidates.

The use of patterns in data to drive scientific discovery has a long history; Mendeleev’s 1869 periodic table is a landmark example, and modern machine learning extends this principle to high-dimensional materials datasets. Building on this perspective, the availability of high-throughput computational resources and curated databases such as SuperCon has enabled the use of machine learning to predict superconducting properties directly from chemical composition [9]. Early successes in this domain demonstrated the feasibility of both classification and regression tasks. Stanev et al. [10] applied a random forest classifier to a few thousand compounds and showed that machine learning can reliably distinguish superconductors from non-superconductors. Hamidieh [11] developed a gradient-boosting model trained on a few thousand known superconductors using elemental descriptors to predict critical temperatures. These efforts marked a shift from manual heuristics to data-driven inference. Subsequent works introduced deep learning architectures tailored to composition-based representations. Konno et al. [12] encoded chemical formulas onto a periodic-table grid and trained convolutional neural networks to recognize superconducting patterns. Pereti et al. [13] employed a DeepSet framework to model unordered element sets and successfully screened minerals for telluride-based superconductors, validating their predictions experimentally. Other approaches have incorpo-

---

\* †Corresponding author: suman.itani@unh.edu,  
yibo.zhang@unh.edu

rated convolutional and recurrent layers [14], or trained deep neural networks directly on electronic band structures to infer phase diagrams [15]. To improve robustness and uncertainty quantification, ensemble methods such as Optuna-optimized stacking [16], variational Bayesian networks, and Monte Carlo dropout have also been explored [17]. Beyond these examples, diverse machine learning models have been applied to superconductivity classification and  $T_c$  prediction [18–28]. These efforts reflect the growing impact of data-driven methods in accelerating superconductor discovery.

Although these data-driven models demonstrate that composition alone encodes useful information, they also reveal significant limitations. Most training sets contain only a few thousand entries and rely on manually engineered descriptors—Magpie features [29], periodic-table images, or other physically motivated quantities—that may bias the learned relationships and limit generalization. Structural information is critical for understanding superconductivity, yet the SuperCon database—the most widely used resource—contains about 16,400 compounds but provides only chemical compositions and critical temperatures; both pressure data and full crystal structures are absent [9]. Models trained solely on composition cannot distinguish structural polymorphs or capture the influence of lattice geometry on  $T_c$ . Efforts to incorporate structure, such as 3DSC [30], S2SNet [31] dataset that matched SuperCon compositions with approximate structures from the Materials Project [32], ICSd [33, 34], improve predictive performance but reduce the number of usable entries because structural data exist for only a subset of compounds. Another limitation is the limited availability of verified non-superconductors. While some negative entries are present, they are far fewer than positive cases, hindering the development of balanced and reliable classifiers.

A major bottleneck in building comprehensive and accurate datasets is that most experimental data reside in unstructured scientific publications and are not readily machine-readable. Manually reading papers to extract structured information is labour-intensive. Rule-based natural-language-processing systems, such as ChemDataExtractor [35, 36], can automate some extraction tasks but require extensive domain expertise and often struggle with complex sentence structures. Recently, large language models (LLMs) trained on massive corpora have shown remarkable capabilities in information extraction [37, 38]. Carefully engineered prompts already enable GPT-4o to parse important structured information for magnetic and thermoelectric compounds with high accuracy [39, 40].

Beyond extraction, fine-tuned versions of GPT-3 and related models can answer chemistry questions, perform property prediction and classify materials phases using only modest amounts of task-specific data [41]. Also, language models have been adapted for generative materials design. AtomGPT encodes chemical formulas and crystal structures as text and fine-tunes transformer ar-

chitectures to predict different materials properties [42]. DiffractGPT uses a generative transformer to reconstruct atomic structures directly from X-ray diffraction patterns [43]. Fine-tuned language models have also been shown to generate chemically valid and stable inorganic compounds with higher success rates than diffusion models [44–46]. Collectively, these studies suggest that, when properly adapted, large language models can function as general-purpose engines for property prediction, structure inference, and even inverse design. However, the application of LLMs to superconductivity—spanning comprehensive data extraction as well as forward and inverse property prediction—remains in its early stages and has thus far been limited to relatively small, curated datasets.

In this work, we use large language models across the entire data-extraction-to-property-prediction pipeline, as shown in Figure 1. First, we extend our previously developed automated text-mining framework [39] with superconductivity-specific prompts to extract structured information directly from the literature. The system identifies relevant papers, parses chemical compositions, records reported transition temperatures along with applied pressures, and captures structural descriptors such as crystal system, space group, and lattice parameters when available. Using this workflow, we constructed a dataset of 78,204 entries, covering about 19,058 unique compositions—including many not present in existing databases. The dataset contains both superconducting and non-superconducting reports and preserves multiple  $T_c$  values for compositions measured under different pressures or doping conditions.

Second, we fine-tuned open-source transformer models (Mistral-7B, Llama-3.1-8B, Qwen3-14B, Qwen3-2507-4B, Phi4-14B) for two supervised tasks: (i) classifying whether a material is superconducting and (ii) predicting its critical temperature. Because these models operate directly on textual representations of chemical formulas and structural metadata, they bypass the need for hand-engineered descriptors. We systematically compared the performance of all fine tuned LLMs across multiple input formats and benchmarked them against feature-based models such as random forests, XGBoost, and fully connected neural networks. The best transformer classifier (Qwen3-14B) achieved an accuracy of 91.1%, essentially matching the performance of the best feature-based baseline (91.0%), while showing stronger generalization to compositions outside the training distribution. For  $T_c$  regression, the best-performing fine-tuned LLMs reached  $R^2 \approx 0.84$ , significantly outperforming their base counterparts and achieving accuracy comparable to feature-based models. We also fine-tuned Qwen3-14B for inverse design, allowing it to generate compositions and space groups conditioned on a target  $T_c$ . In our tests, the model accurately reproduced about 73% of the compositions seen in the training set and generated 27% novel compositions not present in the training data.

Finally, we applied the trained LLM predictors to GeNoME database to identify potential new supercon-

ductors. These models identified – previously unreported materials as potential superconductor candidates. While these predictions have not yet been experimentally validated, they illustrate how combining LLM-based literature extraction with LLM-based prediction provides a scalable framework for accelerating the discovery of superconducting materials.

## II. METHODOLOGY

### A. Automated Database Compilation Using LLMs

We constructed the superconductivity database using an automated pipeline driven by large language models (LLMs), following and extending our earlier approaches [37, 39]. The workflow, shown in Figure 1, begins by identifying relevant scientific papers from sources such as the American Physical Society (APS) and Elsevier. Using keyword searches including “superconductor,” “critical temperature,” “superconducting materials,” and “non-superconductor,” we collected approximately 90,000 unique DOIs.

Many of these articles were available directly in XML format through publisher APIs. These XML files were processed with dedicated parsers to generate structured Markdown documents. However, older APS articles and several handbooks were available only as scanned PDFs containing valuable historical data. These PDF documents were converted to Markdown using specialized text, table, and PDF parsers, with assistance from the Google Gemini LLM.

All Markdown documents were then passed through a set of carefully engineered prompts for advanced LLMs (e.g., GPT-4o, Gemini). The models extracted structured information—including compositions, reported transition temperatures, applied pressures, and available structural metadata—and returned the results in JSON format. These JSON outputs were aggregated to produce the final comprehensive superconductivity database.

### B. Fine-tuning Large Language Models

To adapt foundation models to domain-specific superconductivity prediction tasks, we fine-tuned Mistral-7B [47], LLaMA 3.1-8B [48], Qwen3-14B [49], Qwen3-2507-4B, and Phi4-14B [50] using a parameter-efficient fine-tuning (PEFT) approach based on Low-Rank Adaptation (LoRA) [51], implemented through the Unslloth package [52]. These models were selected because they provide strong general-purpose performance while maintaining compact model sizes, making them feasible to fine-tune on our available GPU resources (a single 24 GB GPU).

In our implementation, the pretrained model was loaded in 4-bit quantization for memory efficiency and

wrapped with LoRA adapters targeting both the self-attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the feed-forward projections (`gate_proj`, `up_proj`, `down_proj`). The LoRA rank and scaling factors were set to  $r = 16$  and  $\alpha = 16$ , respectively, and dropout was disabled for determinism. Gradient checkpointing was enabled to minimize activation memory usage during backpropagation.

For a linear transformation with weight matrix  $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ , LoRA augments it as  $W' = W + \Delta W$ , where  $\Delta W = BA$ , with  $A \in \mathbb{R}^{r \times d_{\text{in}}}$  and  $B \in \mathbb{R}^{d_{\text{out}} \times r}$ , and  $r \ll \min(d_{\text{in}}, d_{\text{out}})$ . During training, only the low-rank matrices  $A$  and  $B$  are updated, leaving the pretrained weights frozen. This approach allows the model to specialize efficiently for materials-science data while preserving the general linguistic and reasoning knowledge learned during large-scale pretraining.

#### 1. Dataset preparation and instruction templates.

All models were fine-tuned using a supervised instruction-tuning paradigm, with training samples formatted in the Alpaca style [53]. Each record contains three fields—`instruction`, `input`, and `output`—which are concatenated into a single training sequence and terminated with an end-of-sequence token ( `eos_token`). The resulting text string is stored under a single field, `text`, for use with the supervised fine-tuning trainer. This unified format is compatible with multiple task types—including binary classification, numerical regression, and inverse design—while each model is fine-tuned separately on the dataset corresponding to its specific task.

*a. Binary classification format.* For the superconductor/non-superconductor classification task, each input consisted only of the chemical composition, and the model was required to output 0 or 1:

```
"instruction": "Given only the chemical
composition, output 1 for superconductor and 0 for
non-superconductor. Output a single character: 0
or 1.", "input":
"Sc0.1625Zr0.1625Nb0.1625Ta0.1625Rh0.175Pd0.175",
"output": "1"
```

We restrict the output vocabulary to numeric tokens (0 or 1) rather than textual labels (e.g., “superconductor”), which avoids token overlap and simplifies the cross-entropy loss.

*b. Composition only-based regression format.* For predicting the superconducting transition temperature ( $T_c$ ) from composition alone, we used:

```
"instruction": "Given only the chemical
composition, predict the superconducting transition
temperature Tc in kelvin (K). Output only a number
in K with no text or units.", "input":
"Y1Ba1.9La0.1Cu407.8", "output": "67.8"
```

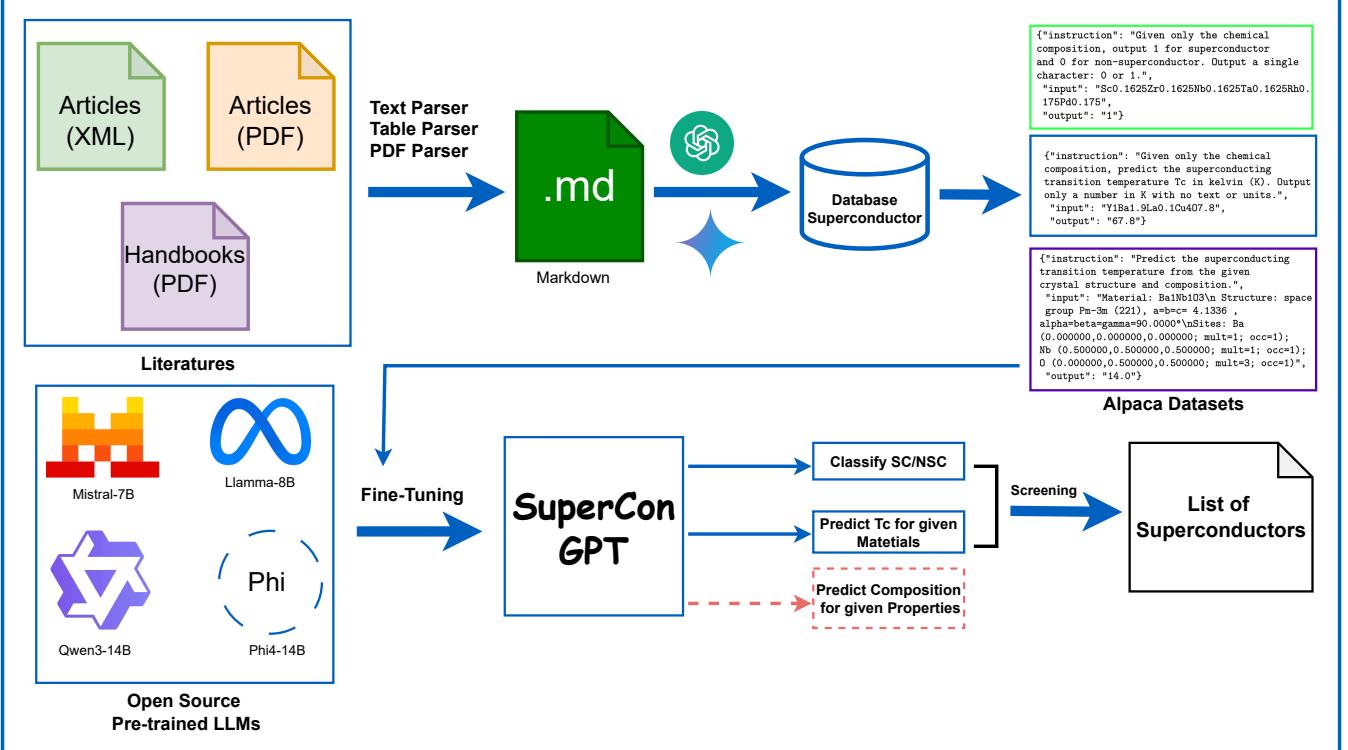


FIG. 1. **End-to-end LLM-driven workflow.** Scientific articles and handbooks are parsed into markdown files, which are processed by an LLM to extract structured data. The resulting database is used to generate Alpaca-format instruction datasets for fine-tuning open-source LLMs (Mistral-7B, Llama 3.1-8B, Qwen3-14B, and Phi-14B). The fine-tuned models perform SC/NSC classification,  $T_c$  regression, and inverse design, enabling large-scale identification of candidate superconductors.

During training, the model minimizes the token-level cross-entropy loss; at evaluation, the output string is parsed as a floating-point value to compute MAE, RMSE, and  $R^2$ .

*c. Regression with crystal system and space group.* To incorporate structural metadata, we designed a second regression dataset that included crystal system, space group, and pressure (when available):

```
"instruction": "Given a material's crystal system, space group, and chemical composition, predict its superconducting transition temperature (Tc).",
"input": "Material: Eu1Fe1.62Co0.38As2 Crystal system: Tetragonal Space group: I4/mmm Pressure: ambient", "output": "5.15 K"
```

*d. Regression with full CIF information.* We additionally constructed a structure-informed dataset using complete crystallographic information files (CIFs). The corresponding template was:

```
"instruction": "Predict the superconducting transition temperature from the given crystal structure and composition.", "input": "Material: Ba1Nb103 Structure: space group Pm-3m (221), a=b=c=4.1336 Å, alpha=beta=gamma=90.0000° Sites: Ba (0.000000,0.000000,0.000000; mult=1; occ=1); Nb (0.500000,0.500000,0.500000; mult=1; occ=1); O (0.000000,0.500000,0.500000; mult=3; occ=1)", "output": "14.0"
```

```
(0.500000,0.500000,0.500000; mult=1; occ=1); O (0.000000,0.500000,0.500000; mult=3; occ=1)",
"output": "14.0"
```

Encoding lattice parameters, atomic positions, and symmetry information in serialized text enables the transformer to attend jointly to structural and compositional descriptors through its self-attention mechanism.

*e. Inverse design format.* Finally, we prepared an inverse design dataset in which the model is asked to propose a superconducting composition along space group, conditioned on a target  $T_c$  and, optionally, a crystal system:

```
"instruction": "Given a target Tc and crystal system, propose a likely superconducting composition.", "input": "Target Tc: 92 K Crystal system: Orthorhombic", "output": "Y0.997Fe0.003Ba2Cu3O7 (space group Pmmm, under ambient pressure); reported Tc = 92.3 K."
```

This dataset allows the model to learn the reverse mapping from desired properties to plausible compositions and structural motifs.

## 2. Training objective and optimization

Fine-tuning was performed using the Unslloth framework [52] together with the `SFTTrainer` module from the Transformers Reinforcement Learning (TRL) library [54]. All models were trained for four epochs using AdamW (8-bit) with a weight decay of 0.01, a learning rate of  $2 \times 10^{-4}$ , and a linear warm-up schedule. Each GPU processed a batch size of 2 with gradient accumulation of 4, giving an effective batch size of 8. Mixed-precision training (fp16/bf16) was employed to reduce memory usage and improve throughput.

During fine-tuning, the model receives tokenized instruction–input–output sequences. Each token is embedded and propagated through a stack of transformer blocks, where information is integrated through multi-head self-attention and position-wise feed-forward layers. For a single attention head, the scaled dot-product attention is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (1)$$

where  $Q$ ,  $K$ , and  $V$  denote the query, key, and value matrices and  $d_k$  is the key dimensionality [55]. Multi-head attention concatenates multiple such heads, enabling the model to attend to different chemical and structural patterns across the serialized sequence.

Supervised fine-tuning optimizes the causal language-modeling cross-entropy loss over the output portion of each training sequence. Given a target token sequence  $y = (y_1, \dots, y_T)$  and model distribution  $p_\theta$ , the loss is

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^T \log p_\theta(y_t | y_{<t}, x), \quad (2)$$

where  $x$  denotes the concatenated instruction and input. During training, only the response tokens contribute to the loss, ensuring that the model learns the desired task mapping (classification, regression, or inverse design).

This instruction-driven formulation enables a unified training setup across all kind of datasets. Parameter-efficient fine-tuning via LoRA adapters allows the transformer to specialize in superconductivity prediction while keeping the base model weights frozen, resulting in domain adaptation with minimal computational overhead.

## C. Feature-based Machine Learning Models

To benchmark the fine-tuned language models, we trained a set of traditional feature-based models for both superconductivity classification and  $T_c$  regression. These models train on manually engineered numerical descriptors. For each composition, we first built an elemental feature vector in which each element is assigned a normalized fractional proportion in the range 0–1 (0 for absent elements, 1 for a pure element). Additional scalar descriptors (e.g., average atomic weight, average atomic

number, average electronegativity, and related compositional statistics) were computed using `pymatgen` library [29, 56].

For the SC/NSC classification task, we trained Random Forest and XGBoost classifiers on the engineered feature set. The data were split into stratified training, validation, and test subsets (overall 80:20 train–test ratio), preserving the class distribution. Hyperparameters were tuned on the validation set, and the final models were evaluated on the held-out test set in terms of accuracy, precision, recall,  $F_1$ -score, and confusion matrices.

For  $T_c$  prediction, we used the same feature construction pipeline. We again applied a stratified train–test split, using quantile bins in  $\log_{10}(T_c + 1)$  to balance different temperature ranges. As baseline regressors, we trained (i) an XGBoost ensemble and (ii) a fully connected feedforward neural network ensemble (iii) random forest. In all cases, multiple models were trained on stratified, down-sampled subsets of the training data, and predictions on the test set were obtained by averaging across ensemble members. Performance was quantified using the coefficient of determination ( $R^2$ ), mean absolute error (MAE), and root mean square error (RMSE), and was used as a reference point for assessing the gains achieved by the fine-tuned LLMs.

## III. RESULTS AND DISCUSSION

### A. Construction of the Experimental Superconductor Database

We developed an experimental superconductor database using the workflow illustrated in Figure 1. A key strength of our approach is that, in addition to compiling superconductors and their associated properties, we also incorporated experimentally verified non-superconductors. This inclusion is critical for training reliable machine learning classifiers capable of distinguishing superconductors from non-superconductors.

The final database contains 78,203 entries, aggregated from sources such as the American Physical Society (APS), Elsevier journals, and domain handbooks on superconductivity. These records correspond to 19,058 unique chemical compositions. For each entry, we extracted the material name, chemical formula, material class (e.g., bulk or thin film), and material type (superconductor or non-superconductor). Critical temperature ( $T_c$ ) values were recorded together with the applied measurement pressure, and, where available, both lower and upper critical magnetic fields were included. We also collected structural descriptors such as crystal system, lattice type, lattice parameters, and space group. Each record explicitly indicates whether the information originates from experimental measurements or theoretical calculations. A complete list of extracted fields is provided in Table I.

The reliability of this workflow is supported by our pre-

vious construction of the Northeast Materials Database (NEMAD), which uses the same LLM-based extraction methodology and was validated through expert review[39]. This prior validation demonstrates the robustness of the approach and provides confidence in the accuracy of the superconductivity database generated here.

A comparison with the widely used SuperCon database shows that our dataset serves as a complementary resource that expands the available experimental information on superconducting materials. Whereas SuperCon primarily reports chemical composition and critical temperature, our database also records structural descriptors, applied pressure, critical fields, and material-class annotations. Among the  $\sim 19,058$  unique compositions in our collection, approximately 14,000 do not appear in SuperCon. These additional compositions and property types provide broader coverage of the superconducting materials landscape and offer a valuable supplementary resource for both superconductivity research and data-driven materials discovery.

The frequency distribution of chemical elements in the database (Figure 2) shows clear trends consistent with the chemistry of known superconductors. Oxygen appears in the largest number of compounds, followed by copper, barium, strontium and other rare-earth or transition-metal elements frequently found in cuprate and perovskite-derived superconductors. This dominance reflects both the historical research emphasis on oxide-based superconductors and the structural tunability of these materials. For clarity, some of these very low-frequency elements are not shown in the figure. Overall, the long-tailed distribution highlights the chemical bias of the literature: a relatively small subset of elements accounts for the majority of experimentally studied superconducting compositions.

The distribution of superconducting transition temperatures (Figure 3) further illustrates this heterogeneity. A strong peak is observed below 20 K, corresponding to the large number of conventional BCS superconductors. A second prominent peak emerges near 90–100 K, dominated by cuprate compounds, while the iron pnictide superconductors form a smaller but distinct peak around 30–40 K. These multimodal trends reflect the fundamental grouping of the superconducting families and show that the dataset captures the full diversity of experimentally explored superconductors. The smoothed density curves also reveal the scarcity of materials with  $T_c$  above  $\sim 150$  K, underscoring the continued difficulty of discovering new high- $T_c$  materials.

## B. Fine-tuning LLMs for Superconductor Classification

Rapid and accurate identification of superconductors remains a fundamental challenge in materials science. Traditional machine learning methods rely on manually

TABLE I. Summary of database features and their types.

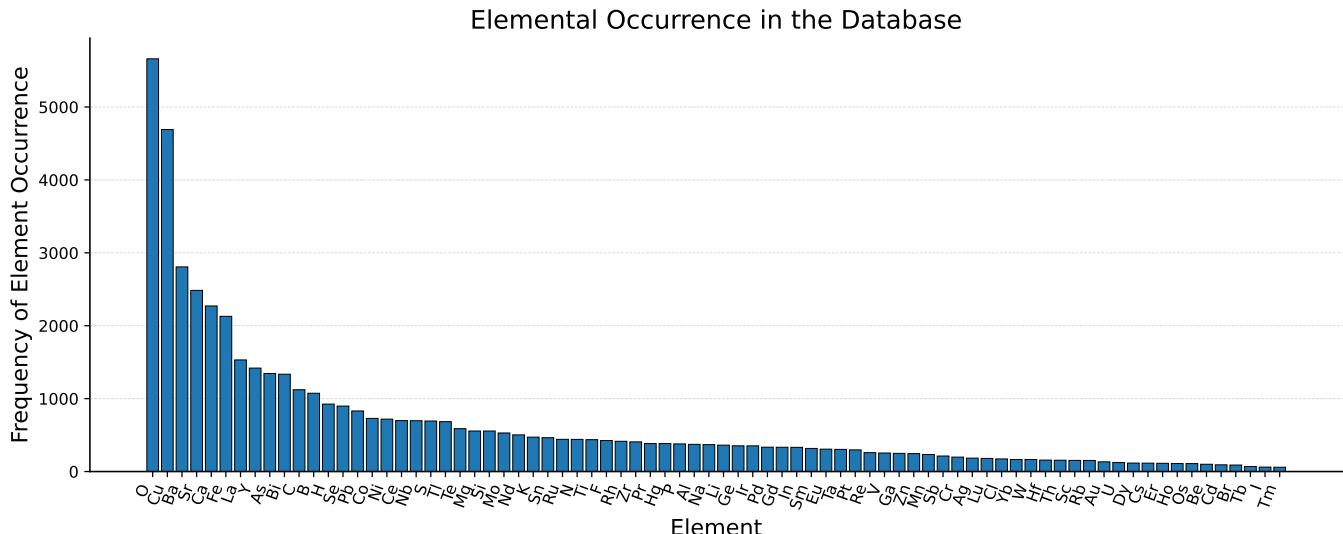
Feature	Type	Unit
Material Name	String	—
Chemical Composition	String	—
Material Class	String	—
Material Type	String	—
Superconductor Type	String	—
$T_c$	Numeric	K
Pressure	Numeric	Pa
Lower Critical Field	Numeric	T
Upper Critical Field	Numeric	T
Crystal Structure	String	—
Lattice Structure	String	—
Lattice Parameters	Numeric	Å
Space Group	String	—
Experimental	Boolean	—
DOI	String	—

designed features, whereas large language models (LLMs) represent a major shift by learning useful representations directly from raw text or chemical compositions through attention mechanisms, removing the need for explicit feature engineering. This approach enables LLMs to infer complex relationships from natural language descriptions or chemical formulas, making them an appealing alternative to conventional classifiers.

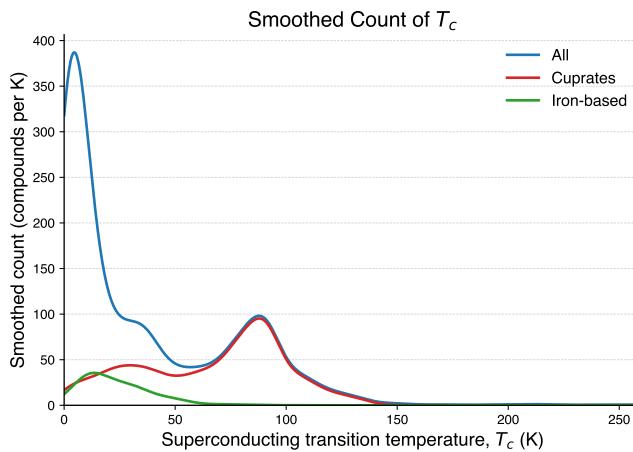
In recent years, numerous open-source LLMs have become readily accessible, providing an unprecedented opportunity to benchmark their capabilities for scientific discovery tasks. However, as our initial tests show, base versions of popular LLMs—such as Mistral-7B, Llama 3.1-8B, Qwen3-14B, Qwen3-2507-4B, and Phi4-14B—are biased toward predicting all compounds as superconductors when queried directly with raw chemical compositions. This bias likely stems from imbalances in the scientific literature: positive (superconductor) results are reported far more frequently than negative (non-superconductor) outcomes, resulting in pre-trained models with limited exposure to the full range of material classes.

To construct a more balanced and representative dataset for supervised fine-tuning, we merged our LLM-generated superconductor database with the SuperCon resource and further supplemented the negative class with experimentally verified non-superconducting magnetic compounds from the NEMAD database, consistent with established practices [12]. This integration resulted in a comprehensive dataset of 45,116 entries, providing a robust foundation for training and evaluation.

Each LLM was then fine-tuned for binary classification (outputting 1 for superconductors and 0 for non-superconductors) using an 80:20 train-test split, with supervised learning implemented via the Unslot package. Importantly, we explicitly instructed the models to output only the class label (as a number) in response to a given chemical composition. This minimizes the genera-



**FIG. 2. Elemental occurrence frequency in the superconductor database.** The bar chart shows the number of distinct compounds in which each element appears, based on the curated superconductor dataset. All elements present in the database are included, with oxygen and copper showing the highest occurrence, reflecting their dominant role in known superconducting materials.



**FIG. 3. Smoothed distribution of superconducting transition temperatures ( $T_c$ ) in the compiled database.** Kernel-smoothed counts are shown for all superconductors (blue), cuprate families (red), and iron-based compounds (green). The distribution exhibits a strong peak at low temperatures and a secondary maximum near 90–100 K associated with cuprate high- $T_c$  materials.

tion of extraneous tokens and ensures that any deviation from the target class is directly penalized in the cross-entropy loss, facilitating clearer optimization and evaluation during training.

For benchmarking, we trained XGBoost and random forest classifiers using 84-dimensional elemental proportion vectors and other composition-based features obtained from the `pymatgen` library [29, 56].

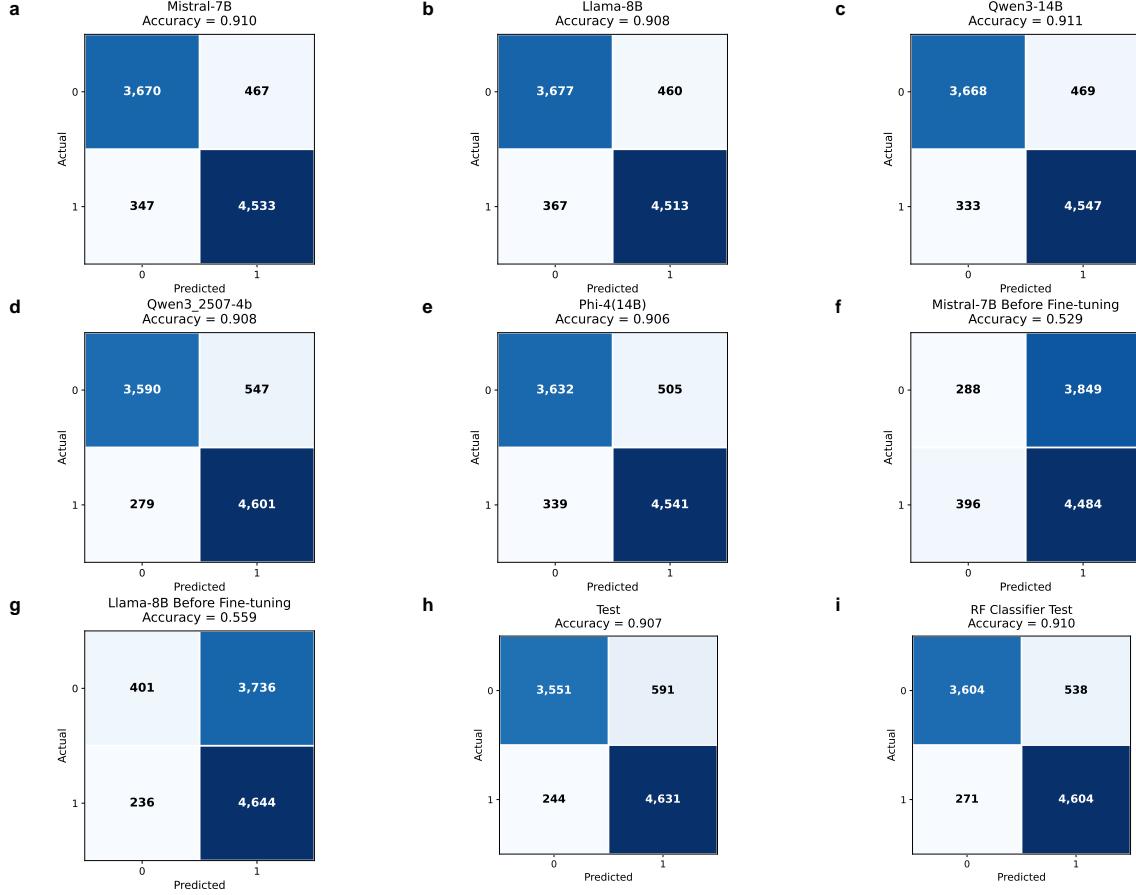
Table II provides a comprehensive comparison of sta-

tistical performance across all models, including precision, recall, F1-score, and accuracy. All fine-tuned LLMs achieve high accuracy, clustering around 91%, with Qwen3-14B delivering the best overall F1-score (0.919) and accuracy (0.911). These results are comparable to, and in some metrics surpass, the performance of established feature-based models.

Figure 4 visualizes the confusion matrices for each approach. Fine-tuned LLMs (Fig. 4a–e) exhibit strong predictive performance, with the majority of predictions aligning along the diagonal and minimal class bias. In contrast, the base models (Fig. 4f, g) display pronounced bias towards the superconductor class, underlining the necessity of targeted fine-tuning. The feature-based models (Fig. 4h, i) also perform well, but do not substantially outperform the fine-tuned LLMs.

Despite these encouraging results, we note that some compounds are consistently misclassified across different models—particularly those with ambiguous or borderline properties. Analyzing these challenging cases may provide further insights into both the limitations of current classification approaches and the underlying complexity of superconductivity as a materials property.

Overall, these findings demonstrate that with appropriate supervised fine-tuning, open-source LLMs can match or exceed the predictive power of traditional feature-based classifiers, while leveraging only raw compositional input. This capability holds significant promise for accelerating the discovery and screening of novel superconducting materials.



**FIG. 4. Confusion matrices for SC/NSC classification using fine-tuned and base LLMs, compared with feature-based classifiers.** (a–e) show the performance of fine-tuned Mistral-7B, Llama-8B, Qwen3-14B, Qwen3-2507-4B, and Phi-14B, all achieving accuracies of approximately 0.90–0.91. (f–g) display the corresponding base (pre-fine-tuning) models, which perform substantially worse (accuracies 0.53–0.56), demonstrating the critical role of supervised instruction tuning. (h) shows the held-out test performance of an XGBoost classifier (accuracy 0.907). (i) provides the Random Forest baseline (accuracy 0.910). Together, these results indicate that fine-tuned LLMs reach classification performance comparable to established feature-based models while significantly improving over their pre-trained counterparts.

**TABLE II. Performance comparison of different classification models.**

Model	Precision	Recall	F1-score	Accuracy
Mistral-7B	0.907	0.929	0.918	0.910
Llama 3.1-8B	0.908	0.925	0.916	0.909
Qwen3-14B	0.907	0.932	<b>0.919</b>	<b>0.911</b>
Qwen3-2507-4B	0.894	0.943	0.918	0.908
Phi4-14B	0.900	0.931	0.915	0.906
Mistral-7B Base	0.538	0.919	0.679	0.529
Llama-8B Base	0.554	<b>0.952</b>	0.701	0.559
XGB Classifier	0.911	0.904	0.906	0.907
RF Classifier	<b>0.913</b>	0.907	0.909	0.910

### C. Fine-tuning LLMs for Predicting the Critical Temperature of Superconductors

Based on the classification model, we can proceed to the prediction of critical temperatures for those classified into superconducting category. Traditional approaches to  $T_c$  prediction—both classical machine learning and deep learning—typically rely on manually engineered features derived from composition or crystal structure. In contrast, large language models (LLMs) can potentially infer complex chemical–physical relationships directly from text or symbolic representations through attention mechanisms, removing the need for handcrafted descriptors.

Here, we fine-tuned a set of open-source transformer-based LLMs—Mistral-7B, Llama 3.1-8B, Qwen3-14B, and Phi4-14B—for  $T_c$  regression tasks. Each model was

trained in a supervised manner using chemical composition as input and the experimental  $T_c$  value as output. To ensure consistency, the models were fine-tuned using the same 80:20 train–test split described earlier. The base models were also evaluated on the same test set for comparison. Figure 5a–d shows the results for the fine-tuned models, whereas Figure 5e–f illustrates the performance of the base versions. The base LLMs exhibited poor regression performance, with  $R^2$  values close to zero or even negative, and high mean absolute errors ( $\text{MAE} \approx 19\text{--}40 \text{ K}$ ), indicating that pre-trained models fail to generalize the quantitative mapping between composition and  $T_c$ . Their predictions cluster heavily around low- $T_c$  regions (typically 10–30 K), suggesting that these models may have primarily encountered BCS-type superconductors during pre-training.

TABLE III. Performance comparison of different models for  $T_c$  prediction.

Model	Dataset	Size	$R^2$	MAE	RMSE
Mistral-7B	C	26k	0.81	7K	15K
Mistral-7B	C+CS+SG	13k	0.84	6K	14K
Mistral-7B	C+Cif	7k	0.73	4K	10K
Qwen3-14B	C	26k	0.81	7K	15K
Qwen3-14B	C+CS+SG	13k	0.83	6K	14K
Qwen3-14B	C+Cif	7k	0.62	4K	11K
Llama 3.1-8B	C	26k	0.78	7K	17K
Llama 3.1-8B	C+CS+SG	13k	0.78	7K	16K
Llama 3.1-8B	C+Cif	7k	0.57	4K	12K
Phi4-14B	C	26k	0.78	7K	16K
Mistral-7B Base	—	—	-1.93	40K	60K
Llama-8B Base	—	—	0.00	19K	35K
XGBoost	C	26k	0.86	6K	12K
Neural Network	C	26k	0.85	6K	13K
Random Forest	C	26k	0.83	8K	14K

After fine-tuning, all models showed substantial improvement: Mistral-7B and Qwen3-14B achieved the highest  $R^2$  values of 0.81, while Llama 3.1-8B and Phi4-14B reached 0.78 (Table III). The corresponding MAE values ranged between 6–7 K, confirming that the fine-tuned models captured both low- and intermediate- $T_c$  regimes with reasonable accuracy. A comprehensive summary of the performance metrics for all models is provided in Table III.

To examine the effect of incorporating structural information, we trained additional models where crystal system (CS), space group (SG), and applied pressure data were appended to the chemical composition (Figure 5g–i). The inclusion of these features led to a modest yet consistent improvement in model performance: for Mistral-7B,  $R^2$  increased from 0.81 to 0.84, and for Qwen3-14B, from 0.81 to 0.83. However, the Llama 3.1-8B model showed no significant gain, maintaining  $R^2 = 0.78$ , suggesting that the benefit of structural metadata depends on the model architecture and how efficiently it integrates auxiliary context.

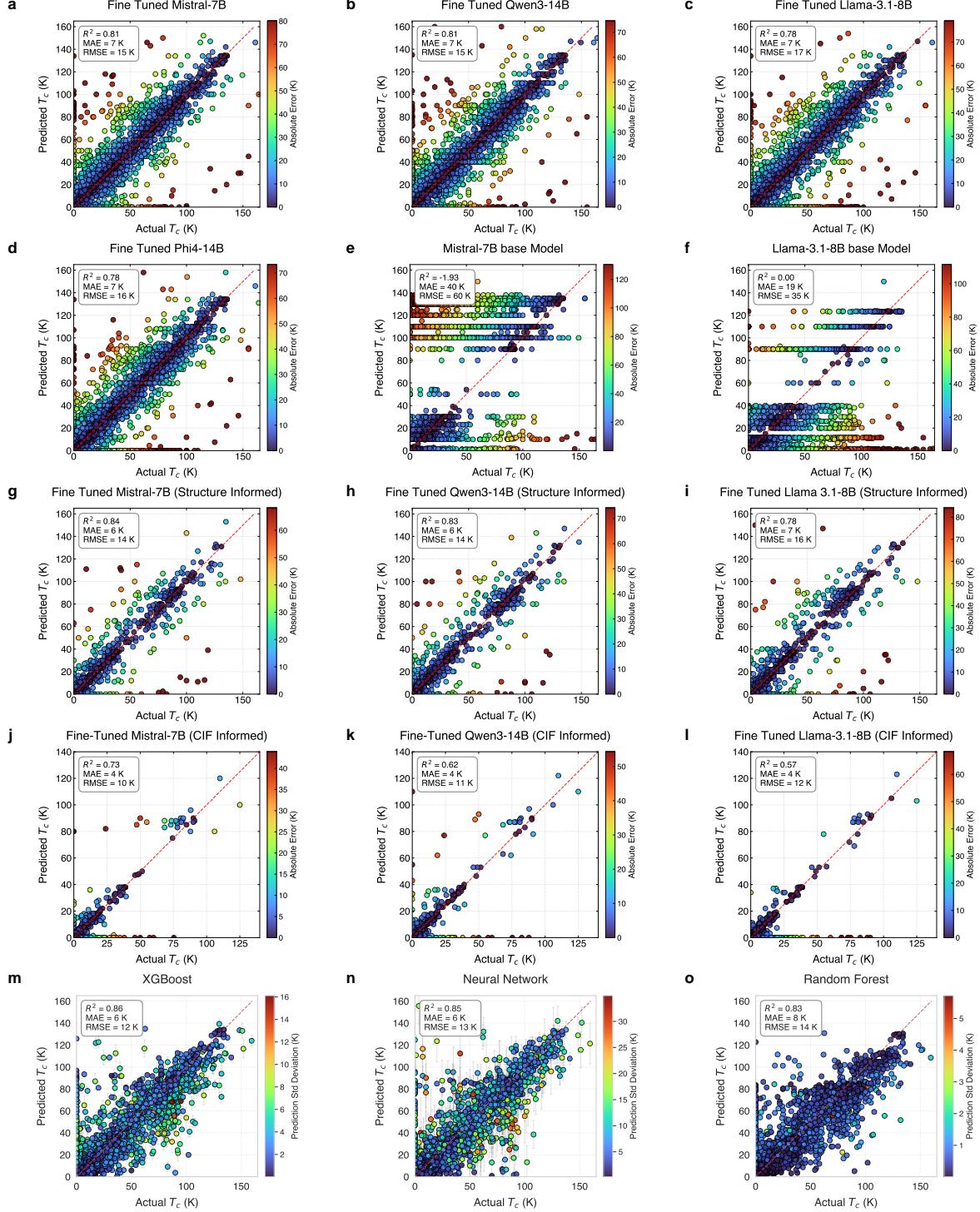
We further extended the dataset by incorporating full

crystallographic information files (CIFs) into the model input. Although our primary database contains only space-group and crystal-system labels, we obtained complete CIFs by matching our compositions and space groups against two external crystallographic resources: the Crystallography Open Database (COD) [57, 58] and the Materials Project [32]. To further increase coverage, we integrated entries from the open-access 3DSC database [30], which provides CIFs for known superconductors. Combining these sources yielded a curated subset of approximately 7,000 materials with full structural information.

Including CIFs significantly increases sequence length and token complexity, since each file encodes lattice parameters, atomic coordinates, Wyckoff positions, and site occupations. The CIF-informed fine-tuned models (Figure 5j–l) achieved  $R^2$  values of 0.73 (Mistral-7B), 0.62 (Qwen3-14B), and 0.57 (LLaMA 3.1-8B), with mean absolute errors remaining low ( $\approx 4 \text{ K}$ ). However, their overall performance decreased compared to the composition-only and CS+SG-informed models. This reduction can be attributed to two factors: (i) CIF inputs produce substantially longer and more complex token sequences, making optimization more challenging, especially under parameter-efficient fine-tuning; and (ii) the CIF-matched subset is considerably smaller and biased toward low- $T_c$  compounds, resulting in reduced data diversity and fewer high-temperature examples for the model to learn from.

For benchmarking, we trained traditional feature-based regression models—XGBoost, neural networks, and random forests—using composition-derived features generated via the `pymatgen` library [29, 56]. As shown in Figure 5m–o and Table III, these models achieved  $R^2$  scores of 0.83–0.86 and MAE of 6–8 K, comparable to or slightly better than the fine-tuned LLMs. Despite this, the fine-tuned LLMs are remarkable in that they learned quantitative structure–property relationships directly from raw chemical formulas, without any feature engineering.

In summary, fine-tuning large language models substantially enhances their capability to predict the critical temperature of superconductors. While current LLM-based approaches achieve accuracy close to feature-based machine learning, their ability to ingest unstructured textual and structural data positions them as powerful, general-purpose tools for data-driven materials discovery. Future optimization—through multimodal fusion of text, composition, and structure representations—may further bridge the gap between language-based and physics-informed prediction models.



**FIG. 5. Performance of fine-tuned LLMs, base models, and feature-based baselines for superconducting transition-temperature ( $T_c$ ) regression. (a–d)** Fine-tuned Mistral-7B, Qwen3-14B, Llama-3.1-8B, and Phi4-14B using composition-only inputs. **(e–f)** Corresponding base (pre-fine-tuning) models evaluated on the same test set, showing substantially lower predictive accuracy. **(g–i)** Models fine-tuned with composition plus crystal system and space group, reflecting improved structure-aware prediction. **(j–l)** Models trained on full CIF-informed inputs, where longer serialized structures and reduced dataset size lead to degraded performance. **(m–o)** Feature-based regressors—XGBoost, a neural-network ensemble, and Random Forest—trained on engineered compositional descriptors.

#### D. Inverse Design of Superconductors Using Fine-tuned LLMs

Beyond forward prediction tasks, a key advantage of large language models (LLMs) is their ability to perform inverse design—generating candidate materials that satisfy desired physical properties. In the context of superconductivity, this task involves proposing chemically and structurally plausible superconducting compositions conditioned on a target critical temperature ( $T_c$ ) and, optionally, a specified crystal system. Achieving reliable inverse design is challenging, particularly due to the scarcity of complete structural information (e.g., CIF files) across known superconductors. As our forward  $T_c$  models already demonstrated limited improvement when full crystallographic inputs were used, generating complete CIFs is currently infeasible with the available data. Therefore, we restricted the inverse design task to predicting chemical compositions and, when possible, their corresponding space groups.

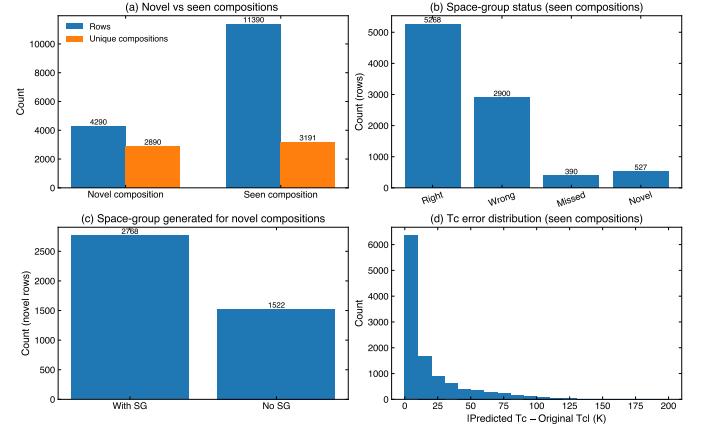
To construct a suitable training set, we generated two types of instruction–response pairs. In the first dataset, the model was given a target  $T_c$  and a crystal system, and was instructed to output a plausible superconducting composition together with its space group and, when applicable, the applied pressure. In the second dataset, the model was conditioned only on a target  $T_c$  and asked to generate a corresponding superconducting composition. Combining both formats resulted in approximately 37,000 instruction–response examples in Alpaca style. Based on its strong performance in both classification and regression, Qwen3-14B was selected as the model for fine-tuning on the inverse design task.

To evaluate the generative performance, we constructed a grid of target conditions consisting of temperatures from 5 to 200 K in 1 K increments and all seven crystal systems, yielding 1,568 unique prompts. For each prompt, we sampled the model ten times to assess generative diversity, resulting in a total of 15,680 model queries. The same instruction formats used during training were applied during inference.

Figure 6 summarizes the performance of the fine-tuned inverse design model. Across all queries, the model produced 4,290 compositions not present in the training set, of which 2,890 were unique (Fig. 6a). This indicates that the model possesses meaningful generative capability and is not merely memorizing its training distribution. Among the novel compositions, 2,768 were generated with both a chemical formula and an associated space group (Fig. 6c), demonstrating that the model can design new materials while simultaneously proposing plausible structural information.

For generated compositions that matched entries in the training set, we evaluated the model’s ability to reproduce the correct space-group information. As shown in Fig. 6b, the model generated the correct space group in 5,268 cases. In other 2,900 cases, however, it assigned a different space group despite the composition being iden-

tical to one in the training set. Although we label these as “wrong”, such differences do not necessarily imply an error, as multiple space groups can be crystallographically feasible for the same composition. In addition, the model omitted the space group entirely for 390 entries even though the training data provided one, while in 527 cases it proposed a novel space group not present in the training set. The latter outcome is noteworthy, as it may reflect the model’s ability to explore alternative symmetry assignments rather than simple memorization. Together, these behaviours illustrate both the capabilities and the current limitations of the model’s structural reasoning.



**FIG. 6. Evaluation of the fine-tuned Qwen3-14B model for inverse design of superconductors.** (a) Number of generated compositions that are novel (not present in the training set) versus those that match compositions present in the training set; among 15,680 model samples, 4,290 generated formulas were novel, of which 2,890 were unique compositions. (b) Space-group accuracy for generated compositions that match entries in the training data. The model reproduced the correct space group in 5,268 cases, assigned an alternative space group in 2,900 cases, omitted the space group in 390 cases, and proposed a previously unseen space group in 527 cases. (c) Space-group generation for novel compositions. Of the 4,290 novel rows, 2,768 included both a chemical formula and a predicted space group, while 1,522 contained only a formula. (d) Distribution of  $|T_c^{\text{target}} - T_c^{\text{original}}|$  for generated compositions that match known materials. Most reproduced compositions lie within 20 K of the target value, indicating that the model preferentially selects materials with experimentally reported  $T_c$  values close to the specified target.

We further evaluated how well the generated compositions satisfy the requested  $T_c$  values for cases where the model reproduced a composition that was already present in the training set. For each such entry, we retrieved its experimental  $T_c$  from the database and computed the absolute deviation  $|T_c^{\text{target}} - T_c^{\text{original}}|$  between the database value and the target value used in the prompt. Figure 6d shows the distribution of this error. The majority of cases lie within 20 K of the requested  $T_c$ , even though the target temperatures were sampled uniformly from 5–200 K.

This indicates that, when it selects compositions from the known data manifold, the model often chooses materials whose known  $T_c$  is close to the specified target.

Although the present inverse design model does not attempt to generate full crystallographic information, it nevertheless serves as a valuable tool for producing synthetic superconducting compositions and enriching existing datasets. As machine learning models for superconductivity continue to advance, such domain-specific synthetic data may help alleviate current data bottlenecks, improve model generalization, and ultimately support the development of more reliable and comprehensive inverse design frameworks.

#### E. Identification of Novel Superconductor Candidates from External Databases

Several large inorganic materials databases—derived from density functional theory calculations or deep-learning-based generative models—contain extensive information on crystal structures and thermodynamic stability but lack any information regarding superconductivity. To explore whether our models can identify previously unrecognized superconducting candidates within these repositories, we screened the GNoME database [59].

To avoid rediscovering known materials, all compositions present in our curated superconductor dataset were first removed from these external datasets. The remaining entries were reformatted to match the input requirements of both our fine-tuned LLM models and the feature-based machine learning models. Screening proceeded in two stages. First, all classification models were applied to label each composition as either superconducting or non-superconducting. Only entries consistently classified as superconductors were retained for further analysis. In the second stage, we applied our regression models—including both fine-tuned LLM regressors and feature-based regressors—to estimate the corresponding critical temperatures.

Through this multi-step screening workflow, we identified a set of previously unreported candidate superconductors with predicted critical temperatures exceeding 10 K for all models. Table IV lists the composition appeared as superconducting by the classification ensemble together with their predicted  $T_c$  values. While these predictions have not yet been experimentally verified, they offer plausible high- $T_c$  candidates emerging from large external materials databases. These compositions may serve as starting points for future experimental exploration or more detailed computational investigations.

## IV. CONCLUSIONS

In this work, we demonstrated an end-to-end framework that integrates large language models into the full pipeline of data extraction, curation, and property pre-

diction for superconducting materials. Using a prompt-engineered extraction workflow, we assembled a comprehensive experimental database of 78,203 records covering 19,058 unique compositions, each annotated with chemical, structural, superconducting transition temperature, and magnetic-field information. The assembled dataset enlarges the available experimental record on superconductors and incorporates property information that is not contained in existing repositories.

We fine-tuned several open-source language models to perform superconductivity classification, critical-temperature regression, and inverse design. The resulting classification models achieve performance comparable to—and in some cases exceeding—traditional feature-based approaches, while also significantly improving upon their base versions. Although feature-engineered models remain the top performers for  $T_c$  regression, the fine-tuned LLMs closely approach their accuracy while substantially outperforming their base versions and learning chemically and structurally meaningful trends directly from text inputs.

Beyond prediction, we showed that fine-tuned models can generate chemically plausible compositions for inverse-design tasks conditioned on a target  $T_c$ . Separately, by applying the forward-prediction models to external materials databases, we identified compounds with predicted transition temperatures above 10 K, demonstrating the utility of LLM-based screening for uncovering new superconducting candidates.

Looking ahead, the framework introduced here can be extended in several ways. Larger-scale language models may further improve regression accuracy, and the incorporation of physics-informed constraints offers a potential route toward more reliable predictions. As broader crystallographic data become available, integrating full structural representations may also enhance both forward prediction and inverse-design performance. These directions highlight opportunities for building increasingly robust, physically grounded LLM-based tools for superconductivity research. A portal on the NEMAD database has been created for the whole community to contribute their recent superconductor data. The database could be a scalable foundation for future data-driven discovery in quantum materials.

## ACKNOWLEDGMENTS

This work was supported by the Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, U.S. Department of Energy, under Award No. DE-SC0020221.

## DATA AVAILABILITY

All resources developed in this work are publicly available. The superconductivity database used

for model training and evaluation is accessible at [www.nemad.org](http://www.nemad.org). All fine-tuned large language models

(LLMs), along with downloadable links and instructions for their use, are provided in a GitHub repository at [github.com/sumanitani/LLM for Superconductors](https://github.com/sumanitani/LLM_for_Superconductors).

## APPENDIX

TABLE IV: Full predicted  $T_c$  (K) values from different models for screened high- $T_c$  materials.

MaterialId	Composition	Space Group	Mistral(C+S)	Mistral(C)	Qwen3(C+S)	Qwen3(C)	XGB	NN
01ed6ada43	Li4NbN3	Ibca	15.0	18.0	10.0	15.3	$20.35 \pm 3.9$	$16.19 \pm 10.72$
05e0b6f484	Na2LiZr(H6Ir)2	P-42m	14.0	14.0	10.0	14.0	$60.09 \pm 8.77$	$24.28 \pm 16.5$
0fe0fc19f1	YZr7H30	R3	115.0	144.0	145.0	144.0	$42.53 \pm 7.36$	$48.49 \pm 28.46$
10d228680c	Ca9AcTaNbN10	P1	10.0	13.0	13.0	14.0	$13.22 \pm 2.08$	$19.44 \pm 15.25$
1245da3d42	Sc5TiH18	C2	135.0	117.0	144.0	101.0	$39.48 \pm 6.38$	$37.87 \pm 15.65$
137be06481	Rb5Ba(MoH9)3	Amm2	117.0	100.0	100.0	100.0	$48.46 \pm 8.44$	$58.08 \pm 37.3$
20ec23fb02	SrLi(H3Pt)2	Fm-3m	11.0	11.0	20.0	14.0	$27.51 \pm 2.9$	$13.77 \pm 11.26$
2abe6657b1	SrLi4	P6.3/mmc	11.0	34.0	11.05	13.5	$11.76 \pm 5.89$	$20.98 \pm 22.11$
2ed395579d	YSc5H18	Cm	145.0	118.0	110.0	146.0	$39.47 \pm 5.89$	$39.67 \pm 13.3$
2f265922e7	Na4Cu2(TcH9)3	Amm2	13.0	11.0	10.05	50.0	$72.66 \pm 15.6$	$49.5 \pm 26.13$
330c818ac4	ThNp6PaH30	R3	11.0	14.3	34.05	14.0	$31.47 \pm 13.71$	$12.82 \pm 12.55$
3519e6cf7a	CaTmMg2(B6Os)2	P2/m	32.0	35.0	33.0	38.0	$25.2 \pm 3.71$	$30.02 \pm 15.86$
390694f010	Li12Ca6Hg4N	Im-3m	30.0	40.0	12.0	12.0	$30.1 \pm 7.07$	$52.35 \pm 27.99$
399df12841	NaB12PC	Ima2	10.0	12.0	10.0	19.0	$21.36 \pm 5.09$	$21.83 \pm 11.4$
3cbabdd56f	NaLi4Ca(RhN)4	P4/m	24.0	39.0	25.0	12.0	$28.35 \pm 6.06$	$23.96 \pm 16.23$
3f9493cc37	KSrH9W	P-62m	140.0	152.0	100.0	146.0	$50.13 \pm 11.55$	$20.67 \pm 30.73$
438be15eb0	Ba5Ca(VH9)3	Amm2	75.0	105.0	100.0	140.0	$51.89 \pm 9.23$	$66.93 \pm 45.57$
4be360c065	ZrSc5H18	C2	110.0	114.0	148.0	145.0	$31.15 \pm 6.11$	$29.52 \pm 11.22$
5050b5f8a4	Ba(Sr2Cu3)2	Amm2	30.0	45.0	60.0	50.0	$55.76 \pm 7.69$	$45.32 \pm 31.25$
544bc02c36	HfMg7(B6Ru)4	Pm	11.0	24.0	25.05	20.0	$23.22 \pm 1.36$	$28.5 \pm 6.86$
56d64c3947	Sc5TaH18	C2	100.0	115.0	108.0	100.0	$26.97 \pm 2.63$	$18.45 \pm 23.34$
57c01b9a23	CaB9N	R-3m	12.0	10.5	10.05	14.5	$20.95 \pm 4.29$	$20.47 \pm 12.45$
5bcdcd48919	Sc5PaH18	C2	110.0	115.0	103.0	143.0	$26.97 \pm 3.21$	$30.33 \pm 22.45$
660ca97914	NaZr(ReH9)2	C2/m	45.0	14.0	11.05	14.0	$73.88 \pm 7.39$	$106.26 \pm 54.09$
6a0c149e22	Mg5Cd(TcH7)2	P-6m2	10.0	16.6	30.05	34.0	$56.31 \pm 5.88$	$34.0 \pm 19.34$
6b825c87e0	KLa2(H6Ir)2	P-3m1	150.0	138.0	138.0	140.0	$37.55 \pm 3.44$	$16.53 \pm 15.92$
6ff2843171	Mg5Cu(TcH7)2	P-6m2	10.0	15.8	30.05	30.0	$54.7 \pm 5.57$	$33.61 \pm 22.27$
710a366bd7	Li3CaAlMo3H20	P1	11.0	30.0	40.0	40.0	$60.63 \pm 11.72$	$83.41 \pm 17.63$
789a40e4d3	Sr6Li12Tl4N	Im-3m	40.0	11.0	10.05	12.0	$37.78 \pm 7.75$	$43.18 \pm 24.12$
7da7d6131f	TmSc5H18	C2	110.0	108.0	103.0	145.0	$37.92 \pm 5.95$	$32.96 \pm 27.27$
805a8fc9a8	KRb(BC7)2	C2	25.0	27.0	10.05	15.0	$24.17 \pm 2.62$	$21.39 \pm 16.28$
84133d870f	Ba2(SrCu2)3	Cmmm	80.0	50.0	80.0	70.0	$58.59 \pm 10.84$	$57.52 \pm 39.22$
896c3355a3	Sr4La2(MoH9)3	Amm2	100.0	65.0	100.0	100.0	$40.15 \pm 10.16$	$37.73 \pm 29.14$
8b3102b39f	RbBaH9W	P-62m	110.0	150.0	100.0	146.0	$46.22 \pm 12.13$	$17.34 \pm 22.79$
8bc068ef0c	SrLaCuN2	I4mm	20.0	40.0	30.0	20.0	$17.77 \pm 3.54$	$31.81 \pm 18.27$
8d727bf7e3	CsAc(BH4)3	P2_12_12	100.0	117.0	100.0	100.0	$67.63 \pm 4.72$	$68.91 \pm 20.88$
9a1ea7b085	NaCa3NbN4	P1	12.0	11.5	15.0	14.0	$19.22 \pm 3.65$	$11.21 \pm 6.97$
9fce4ee521	SrLi3CaNbN4	Pnnm	12.0	32.0	15.0	16.0	$27.35 \pm 8.11$	$20.78 \pm 15.54$
9da8e6e9e0	Mg2TlH8Rh	P4_2/mmc	30.0	80.0	54.0	100.0	$69.36 \pm 3.5$	$71.63 \pm 28.06$
9dffc6eb18	PrMg(ReH9)2	C2/m	70.0	20.0	35.05	14.0	$56.18 \pm 6.97$	$50.68 \pm 35.05$
a0005d43be	H8Se3S	P2_1/m	130.0	110.0	115.0	114.0	$89.04 \pm 10.55$	$86.63 \pm 22.05$
a0da37ef5c	Mg8B23Ru4C	Pm	37.0	37.6	34.0	29.0	$26.14 \pm 0.8$	$32.58 \pm 6.51$
a1f84501d1	YB9N	R-3m	12.0	10.5	10.0	14.0	$18.82 \pm 6.69$	$11.65 \pm 6.74$
a356cab64	Ba6CuC2N7	P2_12_12	40.0	50.0	30.0	55.0	$29.16 \pm 5.86$	$18.31 \pm 20.56$
a454484f40	AcAlH6	R-3m	115.0	100.0	100.0	100.0	$44.51 \pm 5.45$	$58.35 \pm 16.8$
a87ede8637	YMgB6Os	Pbam	11.0	12.0	35.0	29.0	$27.2 \pm 4.18$	$17.2 \pm 8.53$
aa5038f61a	Rb3Ba2BiO6	Fddd	25.0	13.0	11.0	32.0	$51.42 \pm 5.46$	$50.36 \pm 30.1$
ad5ddf6bd9	Th7NpH24	Pm-3	13.9	10.1	30.05	14.4	$17.88 \pm 10.27$	$11.99 \pm 9.65$
b1576dddcc	BPH6	P2_1	115.0	100.0	100.0	104.0	$97.98 \pm 15.95$	$111.11 \pm 17.66$

Continued on next page

TABLE IV: (Continued)

MaterialId	Composition	Space Group	Mistral(C+S)	Mistral(C)	Qwen3(C+S)	Qwen3(C)	XGB	NN
b2d9091623	LiCa2C3F	P4/mmb	12.0	11.5	10.05	11.5	$21.33 \pm 5.4$	$12.67 \pm 9.99$
b3759b4259	LaMg <sub>5</sub> (TcH7)2	P-6m2	30.0	14.0	34.05	47.0	$45.51 \pm 7.81$	$27.59 \pm 21.32$
b850609bf4	CeTh6PaH30	P1	11.2	16.6	11.05	14.5	$19.48 \pm 11.46$	$26.62 \pm 22.52$
bb845e364a	Sr4Ca4Ta(NbN4)3	P1	30.0	35.0	10.05	13.0	$13.55 \pm 4.98$	$15.91 \pm 13.01$
bc769d74fc	Li <sub>6</sub> TcN4	P4.2/nmc	12.0	14.4	25.05	14.5	$26.11 \pm 6.57$	$20.87 \pm 13.19$
be00b8114b	NaLi(B13C2)2	Imm2	50.0	38.0	20.0	30.0	$21.27 \pm 5.33$	$17.9 \pm 9.62$
c3ac0d8b9a	Sr4Mg <sub>2</sub> AlTcH20	P1	37.0	18.0	24.0	20.0	$50.0 \pm 11.19$	$66.33 \pm 30.95$
c4b61bdcce	Mg8B23Os4C	Pm	37.0	38.0	35.0	29.0	$26.26 \pm 0.69$	$25.27 \pm 5.93$
cfea2d0e69	Sr3Ca3(TcH7)2	P3m1	100.0	110.0	60.0	100.0	$63.83 \pm 8.06$	$32.76 \pm 24.87$
d45484b3bc	Th3ScH15	Cc	110.0	109.0	105.0	100.0	$31.21 \pm 7.5$	$30.53 \pm 21.49$
d643242a4b	LiNbH4	P4/mmm	15.0	42.0	33.0	14.0	$43.5 \pm 5.87$	$34.25 \pm 20.59$
de0c587e9e	Ba2Ca7(BN2)6	R-3m	44.0	75.0	50.0	55.0	$15.18 \pm 4.64$	$14.05 \pm 12.69$
dfcc0eade	B(HS)3	P2.1/c	100.0	118.0	100.0	98.0	$64.81 \pm 9.56$	$55.44 \pm 18.57$
e259827901	SrAc3H11	I4mm	115.0	100.0	103.0	100.0	$34.01 \pm 9.04$	$14.66 \pm 18.26$
ea6f316259	H8SeS3	P2.1/m	130.0	110.0	105.0	113.0	$118.11 \pm 12.84$	$109.68 \pm 12.99$
ec9086ecaf	Zr7ScH30	R3	115.0	114.0	155.0	145.0	$30.14 \pm 4.48$	$40.25 \pm 30.17$
ecddef9b73	YMg <sub>5</sub> (TcH7)2	P-6m2	32.0	12.5	34.05	47.0	$49.31 \pm 6.99$	$21.82 \pm 13.99$
eda2f14253	PH3S	P2.1/c	115.0	107.0	100.0	106.0	$103.12 \pm 12.5$	$125.14 \pm 26.17$
f10f4dcd72	Th3B2CN3	Cm	11.0	13.0	25.0	14.0	$17.21 \pm 4.26$	$15.19 \pm 10.6$
f2e3fa7596	BaNbH5	Pnma	12.0	14.0	10.0	14.0	$43.76 \pm 11.07$	$18.28 \pm 15.65$
ffd1c102aa	Li <sub>8</sub> TaNbN6	I2_12_12_1	11.5	39.0	13.0	16.1	$19.29 \pm 2.27$	$18.32 \pm 15.5$
ff16239bad	B13CN	R3m	15.0	13.5	10.0	17.0	$22.31 \pm 5.4$	$32.45 \pm 17.85$
ff75fdc3e2	LiMg <sub>2</sub> AlB28	C2/m	10.0	29.0	20.05	25.0	$20.59 \pm 4.42$	$10.64 \pm 6.61$

- [1] W. V. Hassenzahl, D. W. Hazelton, B. K. Johnson, P. Komarek, M. Noe, and C. T. Reis, Proceedings of the IEEE **92**, 1655 (2004).
- [2] J. M. Gambetta, J. M. Chow, and M. Steffen, npj quantum information **3**, 2 (2017).
- [3] J. G. Bednorz and K. A. Müller, Zeitschrift für Physik B Condensed Matter **64**, 189 (1986).
- [4] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono, Journal of the American Chemical Society **130**, 3296 (2008).
- [5] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. I. Shylin, Nature **525**, 73 (2015).
- [6] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Physical review **108**, 1175 (1957).
- [7] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Physical Review **106**, 162 (1957).
- [8] B. Matthias, in *Progress in low temperature physics*, Vol. 2 (Elsevier, 1957) pp. 138–150.
- [9] M. D. Group, Mdr supercon datasheet ver.220808.
- [10] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, npj Computational Materials **4**, 29 (2018).
- [11] K. Hamidieh, Computational Materials Science **154**, 346 (2018).
- [12] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, and A. Maeda, Physical Review B **103**, 014509 (2021).
- [13] C. Peretti, K. Bernot, T. Guizouarn, F. Laufek, A. Vymazalová, L. Bindi, R. Sessoli, and D. Fanelli, Npj Computational Materials **9**, 71 (2023).
- [14] S. Li, Y. Dan, X. Li, T. Hu, R. Dong, Z. Cao, and J. Hu, Symmetry **12**, 262 (2020).
- [15] J. Li, W. Fang, S. Jin, C. Suo, T. Zhang, Y. Wu, X. Xu, Y. Liu, and D.-X. Yao, AI for Science **1**, 015001 (2025).
- [16] J. Yu, Y. Zhao, R. Pan, X. Zhou, and Z. Wei, ACS omega **8**, 3078 (2023).
- [17] T. D. Le, R. Noumeir, H. L. Quach, J. H. Kim, J. H. Kim, and H. M. Kim, IEEE Transactions on Applied Superconductivity **30**, 1 (2020).
- [18] H. Gashmard, H. Shakeripour, and M. Alaei, Scientific Reports **14**, 3965 (2024).
- [19] D. Kaplan, A. Zheng, J. Blawat, R. Jin, R. J. Cava, V. Oudovenko, G. Kotliar, A. M. Sengupta, and W. Xie, The European Physical Journal Plus **140**, 58 (2025).
- [20] Z. Bai, M. Bhullar, A. Akinpelu, and Y. Yao, Materials Today Physics **43**, 101384 (2024).
- [21] J. Jiang, Y. Xue, L. Zha, S. Yao, B. Wang, W. Hu, L. Peng, T. Shi, J. Chen, X. Liu, et al., Journal of Materials Chemistry C **13**, 9799 (2025).
- [22] S. G. Jung, G. Jung, and J. M. Cole, Journal of Chemical Information and Modeling **64**, 7349 (2024).
- [23] S. Xie, Y. Quan, A. Hire, B. Deng, J. DeStefano, I. Salinas, U. Shah, L. Fanfarillo, J. Lim, J. Kim, et al., npj Computational Materials **8**, 14 (2022).
- [24] O. Lesser, Y. Liu, N. Maus, A. Panigrahi, K. Mallayya, L. M. Schoop, J. R. Gardner, and E.-A. Kim, arXiv preprint arXiv:2510.07373 (2025).
- [25] L. Gu, Y. Liu, P. Chen, H. Huang, N. Chen, Y. Li, T. Lookman, Y. Lu, and Y. Su, Materials Genome Engineering Advances **2**, e48 (2024).
- [26] P. J. G. Nieto, E. G. Gonzalo, L. A. M. García, L. Álvarez-de Prado, and A. B. Sánchez, Alexandria Engineering Journal **86**, 144 (2024).

- [27] E. A. Pogue, A. New, K. McElroy, N. Q. Le, M. J. Pekala, I. McCue, E. Gienger, J. Domenico, E. Hedrick, T. M. McQueen, *et al.*, *npj Computational Materials* **9**, 181 (2023).
- [28] G. Chen, Z. Wang, and F. You, *Journal of Chemical Information and Modeling* **65**, 10871 (2025).
- [29] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, *npj Computational Materials* **2**, 1 (2016).
- [30] T. Sommer, R. Willa, J. Schmalian, and P. Friederich, *Scientific Data* **10**, 816 (2023).
- [31] K. Liu, K. Yang, J. Zhang, and R. Xu, arXiv preprint arXiv:2306.16270 (2023).
- [32] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, *APL materials* **1** (2013).
- [33] G. Bergerhoff, R. Hundt, R. Sievers, and I. Brown, *Journal of chemical information and computer sciences* **23**, 66 (1983).
- [34] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme, *Applied Crystallography* **52**, 918 (2019).
- [35] M. C. Swain and J. M. Cole, *Journal of chemical information and modeling* **56**, 1894 (2016).
- [36] C. J. Court and J. M. Cole, *npj Computational Materials* **6**, 18 (2020).
- [37] Y. Zhang, S. Itani, K. Khanal, E. Okyere, G. Smith, K. Takahashi, and J. Zang, *Journal of Magnetism and Magnetic Materials* **597**, 172001 (2024).
- [38] W. Zhang, Q. Wang, X. Kong, J. Xiong, S. Ni, D. Cao, B. Niu, M. Chen, Y. Li, R. Zhang, *et al.*, *Chemical science* **15**, 10600 (2024).
- [39] S. Itani, Y. Zhang, and J. Zang, *Nature Communications* **16**, 9415 (2025).
- [40] S. Itani, Y. Zhang, and J. Zang, *Computational Materials Science* **253**, 113855 (2025).
- [41] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit, *Nature Machine Intelligence* **6**, 161 (2024).
- [42] K. Choudhary, *The Journal of Physical Chemistry Letters* **15**, 6909 (2024).
- [43] K. Choudhary, *The Journal of Physical Chemistry Letters* **16**, 2110 (2025).
- [44] Z. Cao and L. Wang, arXiv preprint arXiv:2504.02367 (2025).
- [45] N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, and Z. Ulissi, arXiv preprint arXiv:2402.04379 (2024).
- [46] L. M. Antunes, K. T. Butler, and R. Grau-Crespo, *Nature Communications* **15**, 10570 (2024).
- [47] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaud, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, arXiv preprint arXiv:2310.06825 (2023).
- [48] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, arXiv preprint arXiv:2407.21783 (2024).
- [49] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, arXiv preprint arXiv:2505.09388 (2025).
- [50] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, *et al.*, arXiv preprint arXiv:2412.08905 (2024).
- [51] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, *ICLR* **1**, 3 (2022).
- [52] U. AI, D. Han-Chen, and M. Han-Chen, *Unsloth*, <https://github.com/unslothai/unsloth> (2025).
- [53] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, *Stanford alpaca: An instruction-following llama model*, [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca) (2023).
- [54] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Galouédec, *Trl: Transformer reinforcement learning*, <https://github.com/huggingface/trl> (2020).
- [55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Advances in neural information processing systems* **30** (2017).
- [56] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, *Computational Materials Science* **68**, 314 (2013).
- [57] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, *Nucleic Acids Research* **40**, D420 (2012), <https://nar.oxfordjournals.org/content/40/D1/D420.full.pdf+html>.
- [58] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, *Journal of Applied Crystallography* **42**, 726 (2009).
- [59] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, *Nature* 10.1038/s41586-023-06735-9 (2023).