Title: Explainable Synthesizability Prediction of Inorganic Crystal Structures using Large Language Models

Seongmin Kim[1], Joshua Schrier[2*], and Yousung Jung[3,4,5*]

[1] *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Korea*

[2] *Department of Chemistry and Biochemistry, Fordham University, 441 E. Fordham Road, The Bronx, New York 10458, United States*

[3] *Department of Chemical and Biological Engineering (BK21 four), Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea*

[4] *Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea*

[5] *Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korein*

*Email: yousung.jung@snu.ac.kr

*Email: jschrier@fordham.edu

Abstract

We evaluate the ability of machine learning to predict whether a hypothetical crystal structure can be synthesized and explain those predictions to scientists. Fine-tuned large language models (LLMs) trained on a human-readable text description of the target crystal structure perform comparably to previous bespoke convolutional graph neural network methods, but better prediction quality can be achieved by training a positive-unlabeled learning model on a text-embedding representation of the structure. An LLM-based workflow can then be used to generate human-readable explanations for the types of factors governing synthesizability, extract the underlying physical rules, and assess the veracity of those rules. These text-based models can be adapted to specialized cases where less data exists by transfer learning, demonstrated for the case of perovskites.

Introduction

Advancements in computational chemistry and machine learning (ML) have enabled the design and engineering of promising materials with desired properties.[1-10] While the discovery of promising virtual materials has accelerated, the success in experimental validation remains time-consuming.[11,12] To bridge this gap, research has been conducted to limit the exploration to synthesizable materials during the material design process.[13-17] In the field of inorganic materials, thermodynamic energy-based predictions for synthesizability and stability have long been used as crude estimations.[18-23] However, these energy-based predictions often miss many metastable candidate materials and fail to account for materials that are energetically stable yet remain unsynthesized. This indicates that such predictions do not adequately reflect the various complex factors influencing synthesizability. Recently, data-driven approaches based on accumulated synthesized materials have been investigated.[24-29] These

studies have aimed to address the issue using positive-unlabeled (PU) learning, with considering synthesized materials as positive and not-yet-synthesized materials as unlabeled data.[30,31] Notably, in the domain-specific perovskite structures, transfer learning has been effectively employed to achieve accurate synthesizability predictions.[32,33]

However, these data-driven methods have the limitation that the underlying chemical insights used by the machine to predict synthesizability cannot be well understood.[34] Understanding the crucial factors that contribute to synthesizability, rather than simply making predictions, can significantly aid in more feasible materials design. In the field of computer vision, several explainable AI (XAI) techniques have been proposed to understand machine's reasoning for their predictions.[35,36] However, in the field of materials chemistry, such studies are challenging to implement, requiring the need for further research into deep explainability.

Most recently, large language models (LLMs) trained on extensive bodies of literatures have been actively employed to address a variety of chemistry and materials science tasks.[37-45] One powerful approach is to customize pre-trained, general purpose foundation models by fine-tuning them on a small number of examples of a specific task.[46] Recent work has shown that fine-tuned LLMs can achieve performance comparable to existing, complex, bespoke machine learning models for a variety of tasks in organic[47-49] and inorganic[50,51] chemistry, and is the subject of recent comprehensive benchmarking studies.[52,53]

Previously, we showed how fine-tuned LLM could be used to predict inorganic synthesizability and synthesis precursors given only *compositional* information.[50] However, different structures of the same composition (i.e., polymorphs) can have vastly different properties, and in most cases, the goal is to synthesize a particular polymorph. Here, we extend the previous work to consider the synthesizability of specific inorganic crystal *structures*. We compare fine-tuned LLMs which use a text description of the target crystal structure, models constructed using the LLM-embedding representation of that text, and the latest bespoke graph-neural network ML models for this task. We show how to use the LLM to generate explanations of the synthesizability for individual target structures, as well as summarize key factors that generally contribute to synthesizability, and show how to assess the model confidence of those claims. Additionally, we demonstrate the effectiveness of transfer learning techniques, using the specific example of perovskite structures, thereby extending the training and fine-tuning methodologies within the LLM field.

Results and Discussion

General synthesizability prediction

For given general inorganic structural information, the task is to determine whether a structure is synthesizable or not. This is a positive and unlabeled (PU) problem, where we know already-synthesized (positive) and not-yet-synthesized (unlabeled) structures. We closely followed the previous work[24,50] and began with the Materials Project (MP)[54] crystal database retrieved in March 2024, which consists of 60,959 synthesized structures

and 94,402 hypothetical structures. To convert these CIF-formatted structural data into textual data which can be readable as LLM input prompts, we used Robocrystallographer,[55] an open-source toolkit for generating text-based descriptions of crystal structures. Some examples of this conversion are shown in **Figure S1** in the **Supporting Information**. In this work, we used MP30 data (where the number of unique atomic sites in a unit cell is ≤ 30 in the entire MP data) to prevent the text descriptions from becoming too lengthy and exceeding the maximum token limit for LLM input. Similarly, we discarded data where the string length of the text description exceeded 10,000 characters. Accordingly, a total of 100,195 text-described structural data, which consists of 38,347 synthesized and 61,848 hypothetical materials, were prepared, and 20% of the positive and unlabeled data were sampled as a hold-out test dataset for assessing model performance.

We fine-tuned the OpenAI GPT-4o-mini model for the general synthesizability prediction task, following a strategy similar to our previous work.[50] Detailed descriptions of the model, prompt, and fine-tuning process are in the **Supporting Information**. (We also performed parallel experiments using the previous GPT-3.5 base model, but the results were inferior in all cases; see Table S3 and Table S4.) We designed two types of fine-tuned LLM: *StructGPT* is provided with stoichiometric formula information with structural description, and *StoiGPT* contains only stoichiometric information and no structural description. (The general principles of the latter model were described in our recent paper,[50] but here it is retrained and tested on the current dataset with a new GPT base-model.) We compared this to two-types of binary PU-learning classifiers methods: The *PU-CGCNN* model uses a previously graph-based crystal representation;[24] this was retrained with the current dataset. The *PU-GPT-embedding* model first converts the text description of the structure into a 3072-dimensional vector representation using the *text-embedding-3-large* model[56], and then uses that representation as input to train a binary PU-classifier neural network model. The main difference between these two methods is the input representation. Details about model constructions and representations are described in the **Supporting Information**. For model evaluation, only the true positive rate (TPR) or recall can be used as a precisely calculated metric, due to the lack of true negative data in the PU problem. However, the precision (PREC) and the false positive rate (FPR) can be approximated by α-estimation, as discussed in prior works.[57,58] We adopted the same method for model evaluation and comparison in all cases.

As shown in **Figure 1**, the fine-tuned model, StructGPT-FT, outperformed non-fine-tuned GPT model, demonstrating that fine-tuning is crucial for the synthesizability prediction task. (StoiGPT-FT outperformed the StructGPT-FT because a stoichiometry is considered synthesizable if at least one of its various polymorphs has been successfully synthesized, so it is easier to be correct.) StructGPT-FT slightly outperforms the bespoke PU-CGCNN model, indicating that a fine-tuned LLM using the text description of a structure is as powerful as a traditional graph-based learnable representation. This suggests that heuristic decisions in the conventional crystal graph construction, such as limiting edge connection to the 8-12 nearest atoms and omitting geometric angles, insufficiently represent the relevant details of real crystal structures. Even better performance is achieved by combining LLM-based input representation with traditional PU-learning methods. Specifically, the PU-GPT-embedding model outperforms both the StructGPT-FT and PU-CGCNN models, indicating that using a dedicated PU-classifier model is better than using the LLM as a classifier and that GPT-embeddings are more effective than

traditional graph-based representations of structure, respectively. This is our first significant result.

We previously demonstrated the value of fine-tuning to make synthesizability predictions based solely on composition,[50] and the results here demonstrate the added value of including structural information. Recent preprints by Alampara et al.[59] and Song et al.[51] have explored the role of crystal structure descriptions in fine-tuned LLM prediction of solid properties.

Our results support all of these prior claims, and improves upon them by demonstrating the value of using a pre-trained embedding model to generate the representation from the structure as input to a PU-classifier. In addition to the performance benefits, this can also reduce costs. To give a rough approximation, as of Aug 2024, the cost to compute the text-embeddings is $0.065/M tokens (the PU-classifier can be trained and run locally with modest resources which we assume to be free), whereas for the fine-tuning model, the cost is $3/M for fine-tuning and $0.150/M for inference, a saving of 98% and 57%, respectively.

The *text-embedding-3-large* is a hierarchical embedding (also known as *Matryoshka* embedding, by analogy to Russian nested dolls) model, where earlier dimensions correspond to more significant coarse descriptions and later dimensions correspond to increasingly fine-grained features of the text.[56,60] To test whether this is true for structure descriptions, we retrained the PU-GPT-embedding model with inputs that were truncated from the original 3072-dimensions to 2048, 1024, 512 and 256-dimensions. The performance monotonically decreases as the vectors are truncated (Table S8), consistent with the loss of precision. Additionally, while the predicted probabilities are sharply peaked near 0 and 1 for the full vector input, truncating the input causes the distributions to be broadened to intermediate values (Figure S9), indicating that the model is more uncertain about its predictions. Together, these results indicate that the PU-model uses the full vector embedding description to make its prediction. The successful use of these embeddings for prediction suggests their use for determining the similarity of different crystal structures. Whereas previous methods of comparing inorganic crystal similarity have relied primarily upon electronic structure[61,62] or on structural encodings,[63] here the representation comes from a text-description of the structure. This in turn can be used to retrieve similar compounds from a database, which may be useful for LLM-based retrieval augmented generation (RAG) or general discovery by chemical analogy.[64] A full exploration of this is outside the scope of the current article.

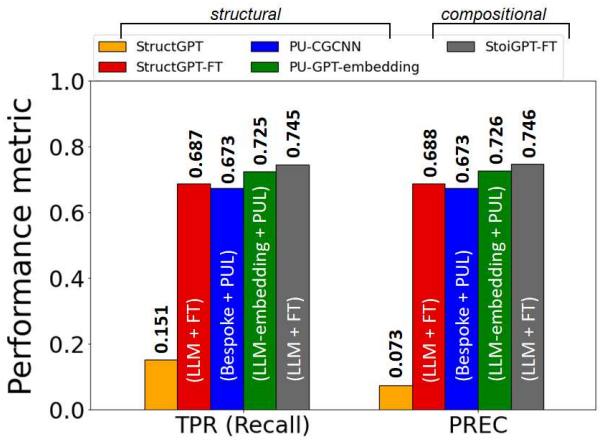


Figure 1. Comparison of model performances for the general synthesizability prediction. FT indicates fine tuning and PUL indicates positive-unlabeled learning. (All calculated metrics are tabulated in Table S1.)

Explanation / Inference for general synthesizability

Using the synthesizability predictions made above, we then used the pre-trained GPT-4o model to generate physical explanations for these results. The user prompt was: "`Explain why an inorganic compound with the following structural information is (not) synthesizable: [Structural description]`". (The "(not)" is included depending on the prediction of our model.) The system prompt was: "`Return only output of the following format for each reason, and no other information: ### Reason 1. **[Keyword of reason]** [Detailed description], ### Reason 2 …`". Using this prompt, we provided GPT-4o with a total of 15,855 structure-prediction pairs, where the predictions of all three models (StructGPT-FT, PU-CGCNN, PU-GPT-embedding) were identical as either positive or negative, to extract the hidden relations and identify the detailed descriptive reasons along with their associated keywords. GPT-4o usually answered the explanation with 4 or 5 reasons (Figure S6). The **Supporting Information** contains examples of these explanations and the URL for the complete set of explanations.

What physical principles does GPT-4o use in these generated reasons? We accumulated all the keywords of reasons for the 15,855 structures, clustered similar words (see the detailed keyword items in the **Supporting Information**), and plotted a bar histogram of the 10 most relevant reasons (**Figure 2a)**. It is generally acknowledged that thermodynamic stability alone is an insufficient factor for material synthesizability. In our results, explanations for about 14% (2,178/15,855) of structures included thermodynamic stability as a reason for their synthesizability. The *text-embedding-3-large* roughly captures the relationships between these categories. We generated the embedding vectors of the detailed description for each type of reason, and plot the results using t-SNE dimensionality reduction in **Figure 2b**, coloring each point by the relevant keyword. The detailed descriptions are separated and clustered by similar keywords, with geometry-related reasons to the right, bond length-related reasons to the bottom, symmetry-related reasons to the top, and thermodynamic /composition

compatibility-related reasons to the left. We can also take this same t-SNE plot and color code each point based on its synthesizability (**Figure 2c).** The explanations are clearly divided according to their predictions across the four partitioned regions. This supports the common-sense notion that all of these factors contribute to synthesizability.

How do these different factors contribute to the synthesizability prediction of our fine-tuned models? We performed an ablation study, using gpt-4o-mini to rewrite the input Robocrystallographer structure description texts, removing specific types of information or to arbitrarily changing specific details (e.g., changing the space group or geometry information). (See section **VI.1 Text elimination and perturbation test** in the **Supporting Information**.) We then provided this modified text as input to the unmodified StructGPT-FT model; results are shown in Table S6. Removing or changing the symmetry and element type information caused the largest degradation in model performance; removing or changing bond-length information had the smallest effect (only reducing the performance by 2-3 percentage points). This is consistent with prior work on structure-based representations in bidirectional LLMs, in which numerical data was entirely omitted from the input text description.[65] Interestingly, while geometry and bond-length are the most commonly invoked reasons by the model, they actually have less impact on the final prediction.

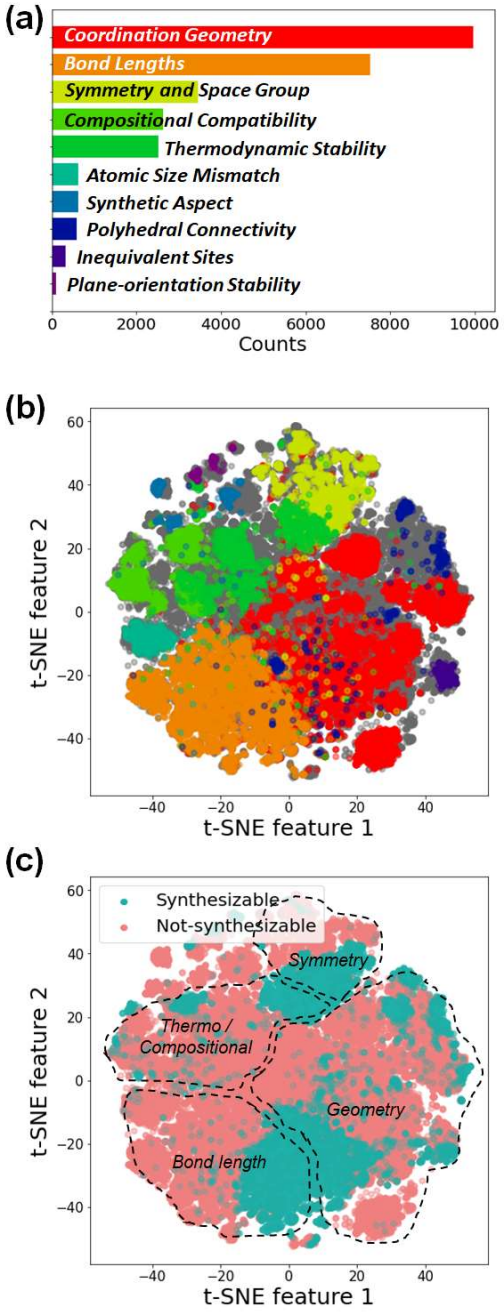


Figure 2. (a) 10 most relevant reasons for general synthesizability. (b) t-SNE for detailed explanation. Each point is colored by its relevant keyword, using the same color scheme as Figure 2a. (c) t-SNE for detailed explanation. Each point is colored by its synthesizability for understanding the reasons across the four partitioned regions.

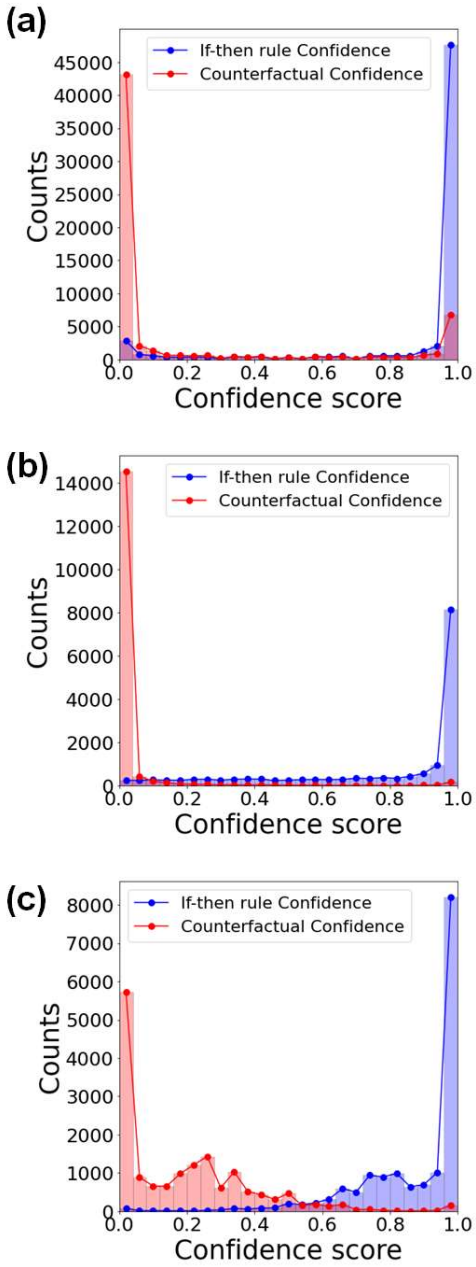


Figure 3. The results of assessing the whole explanations based on the log-probability. Since GPT-4o usually answered the explanation with 4 or 5 reasons per 1 material (Figure S6), we combined (a) each explanation probability by (b) geometric mean of probabilities and (c) arithmetic mean of probabilities.

Are the explanations reasonable? Each generated reason typically consists of a header describing the general type of factor, a sentence describing specific properties of the crystal (copied or paraphrased from the Robocrystallographer input text), and a sentence describing how those specifics relate to stability and synthesizability (or instability and difficulty of synthesis); see examples in the **Supporting Information** Section

V. We developed a four-step approach for extracting the underlying claim and testing the model's confidence in that claim. First, we separate the reasons, removing the header. Second, for all of the sentences that explain the reason, we pass them into a gpt-4o-mini model with the user prompt "`In one sentence, describe an "if-then" rule based on the underlying principle used by this explanation, which could be applied to a new compound: [reason text]`". For example, the input text "The uniform Sc(1)-Ir(1) bond lengths of 2.78 Å indicate a regular and stable bonding environment. Uniform bond lengths generally correspond to lower internal strain in the crystal, suggesting a synthesizable compound." returns "If a new compound exhibits uniform bond lengths, then it likely has low internal strain, indicating that the compound is stable and synthesizable.". We call these outputs *rules*. Third, we pass each rule as a user prompt into gpt-4o, along with the system prompt "`You are provided with a statement of unknown veracity. Return only True or False and nothing else depending on the veracity of the statement.`". The gpt-4o model returns the log-probabilities of each possible response token ("True" or "False"), which allows us to evaluate the probability that the model would answer "True" or "False" (in this case, the model temperature setting is irrelevant). While the associated probabilities of returning "True" or "False" are not strictly the truth of the statement, but they do reflect the model's consensus about the training corpus. Stated another way, a rule for which the model has a high probability of returning "True" is likely to be a principle that is common in the chemistry textbooks and other resources that comprise the training corpus, and thus are a proxy for what the literature would say. Prior work by Kadavath et al. has found that pre-trained LLMs provide well-calibrated true/false self-evaluation on factual questions.[66] Finally, by the classical logical Principle of Non-Contradiction,[67] a statement and its negation cannot both be true. This allows us to generate an internal consistency test, where we have gpt-4o rewrite the original rule by the user prompt "`Rewrite the following sentence so that it would become false: [reason]`". Our above example gets rewritten as "If a new compound exhibits uniform bond lengths, then it likely has high internal strain, indicating that the compound is unstable and unsynthesizable." We then evaluate the veracity as above.

As shown in **Figure 3a**, the probability of being "True" for most if-then rules is close to 1, while for counterfactual rules, the probability of being "True" is close to 0, indicating that most individual explanations are self-evaluated as reasonable by gpt-4o. Since each material has 4 to 5 explanations, to evaluate the veracity of these combined explanations, we calculated the probability distribution which was aggregated using the geometric mean (**Figure 3b**) and the arithmetic mean (**Figure 3c**) of each individual rule. The results also confirmed that the combined explanations for a material exhibit a high degree of reasonability. Furthermore, by calculating the truthiness/falseness confusion matrix between if-then rules and counterfactual rules (Figure S10e), we confirmed that there is internal consistency in gpt-4o's veracity evaluation (when the if-then rule is true, the counterfactual rule becomes false). In this regard, most reasons provided by gpt-4o use principles that are generally well-attested by the training corpus and are internally consistent.

Perovskite synthesizability prediction

To investigate the effectiveness of LLM in a domain-specific material space, we selected perovskite structures as the target space. Given perovskite structural information, the task is to determine whether a structure is synthesizable or not. The data preparation and model construction are similar to those in the **General synthesizability prediction** section. As a result, we constructed a total of 1,533 synthesized and 13,276 hypothetical perovskite structures, and 20% of them were sampled as a hold-out test dataset for assessing model performance. For model construction, we prepared pretrained GPT-4o-mini and fine-tuned versions (StructGPT and StructGPT-FT) for perovskite synthesizability prediction, along with StructGPT-base, which was fine-tuned on general inorganic structures. Since transfer learning has been widely used for solving domain-specific problems in conventional ML field, we adopted it to LLM; we are not aware of previous investigations of this in the chemical LLM fine-tuning literature. Using transfer learning (TL), we refined StructGPT-base on the perovskite dataset (StructGPT-base-FT). For comparison, we also developed PU-GCNN-TL,[32] and PU-GPT-embedding-TL models with transfer learning. All models were evaluated on a hold-out perovskite test dataset. Detailed processes are in the **Supporting Information**.

As shown in **Figure 4**, StructGPT-FT outperformed the non-fine-tuned version, highlighting the importance of fine-tuning for perovskite synthesizability prediction. However, StructGPT-base, fine-tuned only on the general inorganic structures, performed poorly in this domain-specific task. After applying transfer learning, StructGPT-base-FT surpassed StructGPT-FT, showing the effectiveness of TL for LLMs, expanding the training and fine-tuning methodology in the LLM field. Despite this, PU-GCNN-TL still outperformed StructGPT-base-FT, suggesting that dedicated PU-classifier models with TL are more effective for perovskite prediction. Additionally, unlike in the case of **General synthesizability**, PU-GPT-embedding-TL performed slightly worse than PU-GCNN-TL. We speculate that this is related to the relatively simple structure of perovskites. These results are further analyzed in the **Representation capturability** section below.

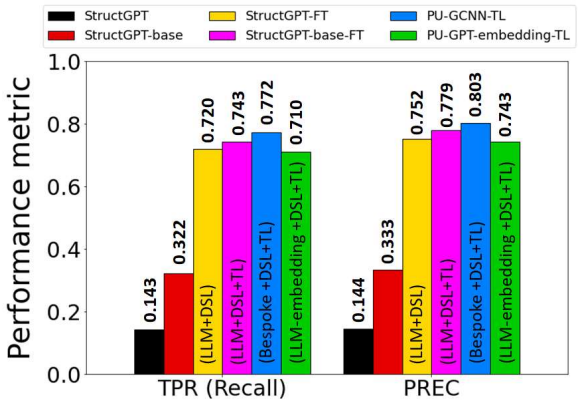


Figure 4. Comparison of model performances for the perovskite synthesizability prediction. DSL indicates domain-specific learning and TL indicates transfer learning. (All calculated metrics are tabulated in Table S2.)

Explanation / Inference for perovskite synthesizability

Perovskite materials have been commonly evaluated for synthesizability using heuristic rules, such as the Goldschmidt tolerance factor,[68,69] which estimates synthesizability based on ionic radii ($t = (r_A + r_B) / (\sqrt{2}(r_B + r_C))$). Traditionally, perovskites are considered synthesizable if $0.7 < t < 1.0$. However, as discussed in previous study,[32] this rule-based screening tends to be overly generous in determining synthesizability in comparison to data-driven machine learning models, since many perovskite materials in the $0.7 < t < 1.0$ range still not being synthesized. We also observed this tendency by comparing this rule to machine learning models (StructGPT-base-FT and PU-GCNN-TL). As shown in **Figure 5a**, the Goldschmidt rule showed the highest recall (98.8%) for 170 hold-out positive perovskites (which satisfy the conditions ABC_3, where C is in [O, Se, F, Cl, Br, and I], B is octahedral site, and ABC_3 is charge neutral compound), whereas two machine learning models showed more modest recall (65.9% and 68.2%). For comparison, the recall for the whole 307 hold-out positive perovskite was 74.3% and 77.2% by StructGPT-base-FT and PU-GCNN-TL, respectively, as shown in **Figure 4**. It also produced a much higher positively predicted ratio among the 1,618 hold-out unlabeled cases (93.2%) compared to the ML models (7.8% and 7.3%), suggesting many potential false positives (**Figure 5b**). This can be attributed to the fact that Goldschmidt screening relies solely on the ionic radii factor with its simplicity.

To better understand these differences, we analyzed 1,331 hold-out hypothetical perovskites predicted as synthesizable (positive) by the Goldschmidt rule but non-synthesizable (negative) by all three ML models. Assuming that the negative predictions for the latter materials are more reasonable predictions since all three machine learning models consistently predict them as negative (while Goldschmidt predicts otherwise), we aim to explore using GPT-4o the explanation of why these materials are not synthesizable. Conducting the same analysis for the latter 1,331 cases as described in the above **Explanation / Inference for general synthesizability** section, GPT-4o analysis (**Figure 5c**) show that the atomic radii and bond related features are indeed the most important aspects in predicting the synthesizability of perovskite, but also additionally suggest that factors beyond ionic radii—such as coordination geometry, compositional compatibility, and thermodynamic stability—are crucial in determining synthesizability. These analyses suggest possible directions to improve the Goldschmidt rule in future works.

To assess the explanation veracity, we also conducted same evaluation manner, as discussed in the above **Explanation / Inference for general synthesizability** section. We provided GPT-4o with a total of 2,679 perovskite structure-prediction pairs, where the predictions of all three models (StructGPT-base-FT, PU-GCNN-TL, PU-GPT-embedding-TL) were identical as either positive or negative to extract their explanations. The results were also confirmed that the explanations for a perovskite material exhibit a high degree of reasonability and their internal consistency (Figure S11 in the **Supporting Information**).

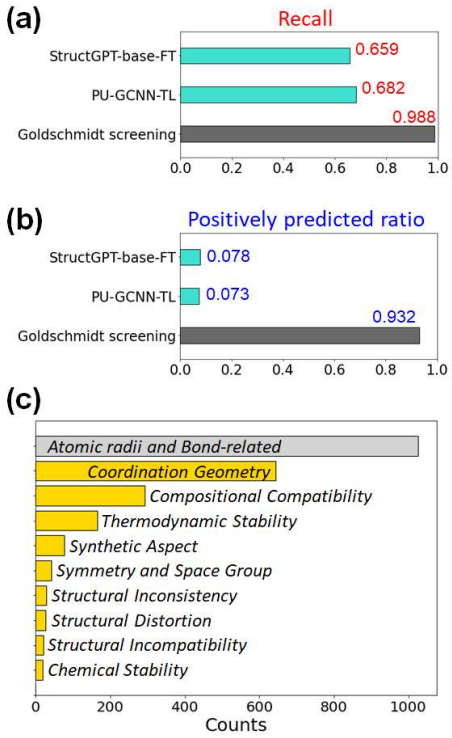


Figure 5. (a) Recalls and (b) positively predicted ratios of the Goldschmidt-based screening and the machine learning models (StructGPT-base-FT and PU-GCNN-TL) for (a) 170 hold-out positive and (b) 1,618 hold-out unlabeled perovskites (ABC_3; C is in [O, Se, F, Cl, Br, and I], B is octahedral site, and ABC_3 is charge neutral compound). (c) 10 most relevant not-synthesizable reasons for 1,331 false positive perovskite cases (synthesizable by the Goldschmidt-based screening but not-synthesizable by all three machine learning models).

Representation capturability

As shown in **Figure 1** and **Figure 4**, the two bespoke ML models, PU-GPT-embedding and PU-CGCNN, exhibited different performance superiority. For the general synthesizability, GPT-embedding showed better representation capturability than the graph-based one, whereas for the domain-specific perovskite synthesizability, the opposite trend was observed. To investigate these contrasting results, we plotted the distribution of text description length (number of characters) for general inorganic crystal structures and perovskite structures, as shown in **Figure 6a**. As depicted in **Figure 6a**, perovskite structures mainly consist of much shorter description lengths due to their simpler formula (ABX_3) and high structural symmetry. In contrast, general inorganic crystals exhibit diverse structures, resulting in a wider range of description lengths. This data discrepancy between general inorganic and perovskite structures contributed to the opposite performance trend.

Based on this observation, we further evaluated both the PU-GPT-embedding and PU-CGCNN models for the general synthesizability prediction across different description length divisions (<1000, 2000, 3000, …, 10000) of the hold-out-test dataset, as shown in **Figure 6b**. The result demonstrated that the representation

capturability of the two models varied as description length increases. This suggests that the representation capturability for inorganic structures can be influenced by structural complexity. For simpler structures, graph-based representation well captures the relations, whereas for more complex structures, GPT-embedding vectors are more effective structural representations. Therefore, we suggest that the appropriate structural representation in model development should be carefully chosen to enhance the model performance depending on the structural complexity of the target system.

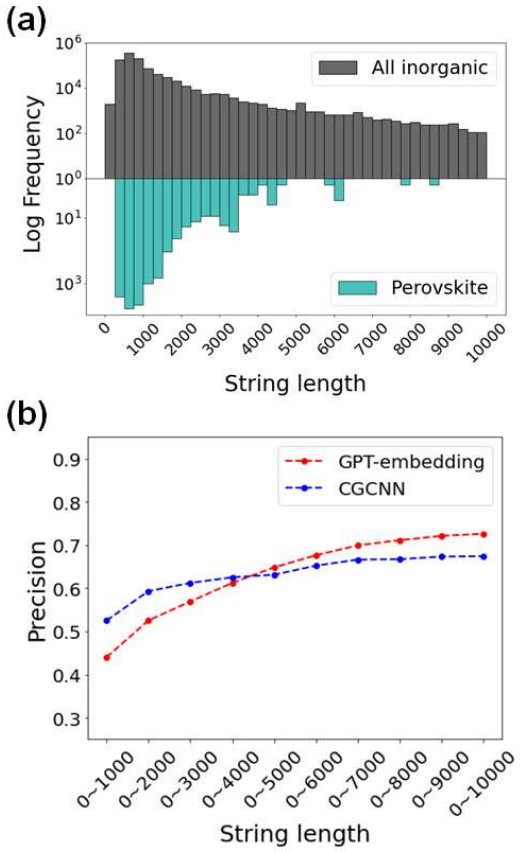


Figure 6. (a) Paired histogram of structural description length for the general inorganic crystal structures and the perovskite structures. (b) Model performances of PU-GPT-embedding and PU-CGCNN depending on the description length divisions of the hold-out-test dataset.

Conclusion

We utilized LLMs for structure-based general synthesizability prediction and domain-specific perovskite synthesizability prediction, along with their explanations. Fine-tuned LLMs and LLM-embedding-based bespoke ML models showed promising performance compared to the traditional bespoke ML models. Furthermore, LLMs can provide explainability by inferring the reasons for determining the synthesizability. Explanations can be easily obtained through a simple prompt. Unlike recent work on using LLMs for materials structure-property explainability,[70] these explanations are applied to model predictions, rather than requiring literature examples.

Based on these explanations, we can specify the detailed and essential aspects related to general synthesizability determination. By employing this strategy for non-synthesizable materials, we can identify the factors contributing to their low synthesizability. We anticipate that these explanations can guide chemists in modifying or optimizing non-synthesizable hypothetical structures to make them synthesizable.

To develop an effective LLM for the synthesizability of perovskite structures, we adopted transfer learning technique to LLMs. The result showed that transfer learning can be effectively applied to LLMs, broadening the training and fine-tuning methodologies within the LLM field. LLM explanations for determining perovskite synthesizability showed some relativeness to the conventional heuristic rule-based screening but also encompassed much broader factors, reflecting the complexity of materials chemistry.

In comparing the graph-based model with the LLM-embedding-based model, we analyzed the representation capturability for inorganic crystal structures. The result showed that LLM-embedding vectors can serve as a more effective structural representation in the case of complex structures compared to the conventional graph-based formulation.

However, there are limitations that should be addressed in future works:

(1) Our prediction-explanation strategy is a disjointed approach, unlike most other explainable AI (XAI) methods[35,36] that are based on input, output, and the model. That is, we used a baseline LLM to infer explanation for fine-tuned LLM. As a result, the explanation might be derived from the general chemical knowledge that were used to pretrain the baseline LLM.

(2) Since this model highly relies on existing material database, the prediction (even the transfer-learned model) could be biased by the distribution of already-synthesized materials.[71-73] Other results suggest that LLM-based methods may be less transferable to out-of-domain problems than conventional methods, due to the lack of hard-coded inductive biases[74]

(3) In this study, we focused on inorganic crystal structures that are well-crystallized and did not consider defects or disordering. However, various levels of defects and disordering often occur in real world materials,[75] indicating the need for future research that can address these aspects as well.

(4) We used the off-the-shelf embeddings, pre-trained Matryoshka embeddings, without additional training or fine-tuning embedding models. However, there is still potential to fine-tune the embedding model using sentence transformers for RAG to explore more advanced LLM embeddings.[76,77] Alternatively, introducing material-specific latent vectors through unsupervised learning of text taken from abstracts in the materials science literature,[78,79] Robocrystallographer structure descriptions,[80,81] or directly on the text of CIF files[82] could be further approaches.

As our goal was to propose the approach of leveraging LLMs for predicting structure-based synthesizability and inferring its chemical explanation, there are many possible ways to improve the performance. In the future, more advanced LLMs can be utilized for developing fine-tuned LLMs, as we demonstrated through the performance comparison between fine-tuned GPT-3.5 and fine-tuned GPT-4o-mini. Designing detailed prompts or combining external functional tools could also contribute to further development. Finally, we hope that ongoing rapid advancements in LLMs will enhance performance.

Acknowledgements

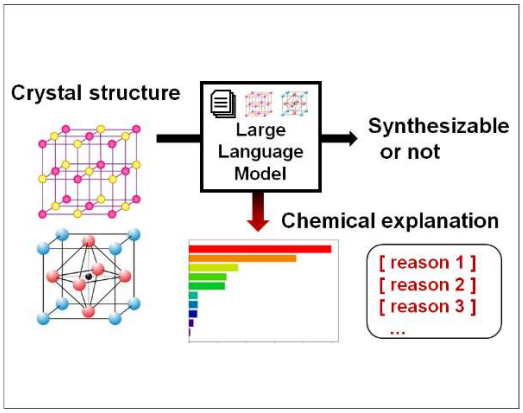
Y.J. acknowledges support from NRF (RS-2023-00283902, 2021R1A5A1030054) and IITP (RS-2021-II211343) of Korea government. J.S. acknowledges Fordham University for granting a sabbatical leave, Seoul National University for a Global Visiting Faculty Fellowship during which the work was initiated, and support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Heavy Element Chemistry Program under contract KC0302031, subcontracted through Los Alamos National Laboratory.

Code availability

The code and data underlying this study are openly available on Github at https://github.com/snu-micc/StructLLM/, with a persistent archival copy deposited at [ZENODO]. Access to GPT-4o-mini and GPT-4o is commercially available to the public at https://openai.com. The PU-CGCNN source code is available at https://github.com/snu-micc/Synthesizability-PU-CGCNN/ and the PU-GCNN-TL source code is available at https://github.com/kaist-amsg/PerovskiteSynthesizability_Manuscript2021/.

Keywords: large language models • inorganic • synthesizability • explainability • crystal representation **(5 maximum)**

Table of Contents



References

1 Pulido, A. *et al.* Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657-664 (2017).

2 Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nature chemistry* **1**, 37-46 (2009).

3 Zunger, A. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry* **2**, 0121 (2018).

4 Jansen, M. Conceptual inorganic materials discovery–a road map. *Advanced Materials* **27**, 3229-3242 (2015).

5 Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature materials* **12**, 191-201 (2013).

6 Muy, S. *et al.* High-throughput screening of solid-state Li-ion conductors using lattice-dynamics descriptors. . *iScience* **16**, 270-282 (2019).

7 Zhuo, Y., Mansouri Tehrani, A., Oliynyk, A. O., Duke, A. C. & Brgoch, J. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nature communications* **9**, 4377 (2018).

8 Zhou, J. *et al.* Discovery of hidden classes of layered electrides by extensive high-throughput material screening. *Chemistry of Materials* **31**, 1860-1868 (2019).

9 Mueller, T., Hautier, G., Jain, A. & Ceder, G. Evaluation of tavorite-structured cathode materials for lithium-ion batteries using high-throughput computing. *Chemistry of materials* **23**, 3854-3862 (2011).

10 Mok, D. H. & Back, S. Generative Language Model for Catalyst Discovery. *arXiv preprint arXiv:2407.14040* (2024).

11 Yano, J. *et al.* The case for data science in experimental chemistry: examples and recommendations. *Nature Reviews Chemistry* **6**, 357-370 (2022).

12 Back, S. *et al.* Accelerated chemical science with AI. *Digital Discovery* **3**, 23-33 (2024).

13 Singh, A. K., Montoya, J. H., Gregoire, J. M. & Persson, K. A. Robust and synthesizable photocatalysts for CO2 reduction: a data-driven materials discovery. *Nature communications* **10**, 443 (2019).

14 Noh, J. *et al.* Path-aware and structure-preserving generation of synthetically accessible molecules. *International Conference on Machine Learning*, 16952-16968 (2022).

15 Chen, S. & Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *Journal of Cheminformatics* **16**, 83 (2024).

16 Park, H., Onwuli, A., Butler, K. & Walsh, A. Mapping inorganic crystal chemical space. *Faraday Discussions* (2024). https://doi.org:10.1039/D4FD00063C

17 Gruver, N. *et al.* Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379* (2024).

18 Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Science advances* **2**, e1600225 (2016).

19 Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Science advances* **4**, eaaq0148 (2018).

20 Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *Journal of the American Chemical Society* **143**, 9244-9259 (2021).

21 Ye, W., Chen, C., Wang, Z., Chu, I.-H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nature communications* **9**, 3800 (2018).

22 Bartel, C. J. *et al.* A critical examination of compound stability predictions from machine-learned formation energies. *npj computational materials* **6**, 97 (2020).

23 Bartel, C. J., Weimer, A. W., Lany, S., Musgrave, C. B. & Holder, A. M. The role of decomposition reactions in assessing first-principles predictions of solid stability. *npj Computational Materials* **5**, 4 (2019).

24 Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-based synthesizability prediction of crystals using partially supervised learning. *Journal of the American Chemical Society* **142**, 18836-18843 (2020).

25 Frey, N. C. *et al.* Prediction of synthesis of 2D metal carbides and nitrides (MXenes) and their precursors with positive and unlabeled machine learning. *ACS nano* **13**, 3031-3041 (2019).

26 Jang, J. *et al.* Synthesizability of materials stoichiometry using semi-supervised learning. *Matter* **7**, 2294-2312 (2024).

27 Antoniuk, E. R. *et al.* Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions. *npj Computational Materials* **9**, 155 (2023).

28 Zhu, R. *et al.* Predicting synthesizability using machine learning on databases of existing inorganic materials. *ACS omega* **8**, 8210-8218 (2023).

29 Davariashtiyani, A., Kadkhodaie, Z. & Kadkhodaei, S. Predicting synthesizability of crystalline materials via deep learning. *Communications Materials* **2**, 115 (2021).

30 Mordelet, F. & Vert, J.-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters* **37**, 201-209 (2014).

31 Bekker, J. & Davis, J. Learning from positive and unlabeled data: A survey. *Machine Learning* **109**, 719-760 (2020).

32 Gu, G. H., Jang, J., Noh, J., Walsh, A. & Jung, Y. Perovskite synthesizability using graph neural networks. *npj Computational Materials* **8**, 71 (2022).

33 Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345-1359 (2009).

34 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547-555 (2018).

35 Gunning, D. *et al.* XAI—Explainable artificial intelligence. *Science robotics* **4**, eaay7120 (2019).

36 Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition* **65**, 211-222 (2017).

37 Anstine, D. M. & Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society* **145**, 8736-8750 (2023).

38 M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1-11 (2024).

39 Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570-578 (2023).

40 Ramos, M. C., Collison, C. J. & White, A. D. A Review of Large Language Models and Autonomous Agents in Chemistry. *arXiv preprint arXiv:2407.01603* (2024).

41 Lei, G., Docherty, R. & Cooper, S. J. Materials science in the era of large language models: a perspective. *Digital Discovery* **3**, 1257-1272 (2024).

42 Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society* **145**, 18048-18062 (2023).

43 Schrier, J. Comment on "Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases". *Journal of Chemical Education* **101**, 1782-1784 (2024).

44 Mirza, A. *et al.* Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475* (2024).

45 Wang, H. *et al.* Evaluating the Performance and Robustness of LLMs in Materials Science Q&A and Property Predictions. *arXiv preprint arXiv:2409.14572* (2024).

46 Dinh, T. *et al.* Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems* **35**, 11763-11784 (2022).

47 Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **6**, 161-169 (2024).

48 Xie, Z. *et al.* Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chemical science* **15**, 500-510 (2024).

49 Zhong, S. & Guan, X. Developing Quantitative Structure–Activity Relationship (QSAR) Models for Water Contaminants' Activities/Properties by Fine-Tuning GPT-3 Models. *Environmental Science & Technology Letters* **10**, 872-877 (2023).

50 Kim, S., Jung, Y. & Schrier, J. Large Language Models for Inorganic Synthesis Predictions. *Journal of the American Chemical Society* **146**, 19654-19659 (2024).

51 Song, Z., Lu, S., Ju, M., Zhou, Q. & Wang, J. Is Large Language Model All You Need to Predict the Synthesizability and Precursors of Crystal Structures? *arXiv preprint arXiv:2407.07016* (2024).

52 Jacobs, R. *et al.* Regression with Large Language Models for Materials and Molecular Property Prediction. *arXiv preprint arXiv:2409.06080* (2024).
53 van Herck, J. *et al.* Assessment of Fine-Tuned Large Language Models for Real-World Chemistry and Material Science Applications. *ChemRxiv preprint chemrxiv:2024.mm31v* (2024).
54 Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **1**, 011002 (2013).
55 Ganose, A. M. & Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications* **9**, 874-881 (2019).
56 OpenAI embedding model Service. https://openai.com/index/new-embedding-models-and-api-updates (2024).
57 Jain, S., White, M. & Radivojac, P. Recovering true classifier performance in positive-unlabeled learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **31** (2017).
58 Zeiberg, D., Jain, S. & Radivojac, P. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 6729-6736 (2020).
59 Alampara, N., Miret, S. & Jablonka, K. M. MatText: Do Language Models Need More than Text & Scale for Materials Modeling? *arXiv preprint arXiv:2406.17295* (2024).
60 Kusupati, A. *et al.* Matryoshka representation learning. *Advances in Neural Information Processing Systems* **35**, 30233-30249 (2022).
61 Isayev, O. *et al.* Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials* **27**, 735-743 (2015).
62 Kuban, M. *et al.* Similarity of materials and data-quality assessment by fingerprinting. *MRS Bulletin* **47**, 991-999 (2022).
63 Li, S. *et al.* Encoding the atomic structure for machine learning in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **12**, e1558 (2022).
64 Rouvray, D. H. Similarity in chemistry: past, present and future. *Molecular Similarity I*, 1-30 (2005).
65 Rubungo, A. N., Arnold, C., Rand, B. P. & Dieng, A. B. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029* (2023).
66 Kadavath, S. *et al.* Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).
67 Horn, L. R. Contradiction. The Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/archives/fall2024/entries/contradiction/ (2024).
68 Goldschmidt, V. M. Die gesetze der krystallochemie. *Naturwissenschaften* **14**, 477-485 (1926).
69 Bartel, C. J. *et al.* New tolerance factor to predict the stability of perovskite oxides and

halides. *Science advances* **5**, eaav0693 (2019).

70 Liu, Q. *et al.* Beyond designer's knowledge: Generating materials design hypotheses via large language models. *arXiv preprint arXiv:2409.06756* (2024).

71 Sun, W. & David, N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discussions* (2024). https://doi.org:10.1039/D4FD00112E

72 Jia, X. *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251-255 (2019).

73 Schrier, J., Norquist, A. J., Buonassisi, T. & Brgoch, J. In pursuit of the exceptional: Research directions for machine learning in chemical and materials science. *Journal of the American Chemical Society* **145**, 21699-21716 (2023).

74 Li, K. *et al.* Probing out-of-distribution generalization in machine learning for materials. *arXiv preprint arXiv:2406.06489* (2024).

75 Cheetham, A. K. & Seshadri, R. Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. *Chemistry of Materials* **36**, 3490-3495 (2024).

76 Training and Finetuning Embedding Models with Sentence Transformers v3. *https://huggingface.co/blog/train-sentence-transformers* (2024).

77 Fine-tune Embedding models for Retrieval Augmented Generation (RAG). *https://www.philschmid.de/fine-tune-embedding-model-for-rag* (2024).

78 Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95-98 (2019).

79 Zhang, B. *et al.* Label-Free Data Mining of Scientific Literature by Unsupervised Syntactic Distance Analysis. *The Journal of Physical Chemistry Letters* **15**, 212-219 (2023).

80 Sayeed, H. M., Baird, S. G. & Sparks, T. D. Structure feature vectors derived from Robocrystallographer text descriptions of crystal structures using word embeddings. *ChemRxiv preprint chemrxiv:2023.3q8wj* (2023).

81 Qu, J. *et al.* Leveraging language representation for materials exploration and discovery. *npj Computational Materials* **10**, 58 (2024).

82 Yadav, L. Atoms as words: A novel approach to deciphering material properties using NLP-inspired machine learning on crystallographic information files (CIFs). *AIP Advances* **14** (2024).