Explainable Synthesizability Prediction of Inorganic Crystal Polymorphs using Large Language Models

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

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

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

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

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

[5] *Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea*

*Email: jschrier@fordham.edu

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

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 explanations can guide chemists in modifying or optimizing non-synthesizable hypothetical structures to make them more feasible for materials design.

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.[24] 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.[25-30] 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.[31,32] Notably, in the domain-specific perovskite structures, transfer learning has been effectively employed to achieve accurate synthesizability predictions.[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.[37-46] 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.[47] 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[48-50] and inorganic[51,52] chemistry, and is the subject of recent comprehensive benchmarking studies.[53-55]

Previously, we showed how fine-tuned LLM could be used to predict inorganic synthesizability and synthesis precursors given only compositional information.[51] However, different structures of the same composition can have vastly different properties, and in most cases, the goal is to synthesize a particular polymorph.

Here, we show that the fine-tuned LLM based on text descriptions of the target crystal structure can give synthesizability predictive performance comparable to the latest bespoke graph-neural network ML models. Moreover, even better prediction performance can be obtained by training a neural network model on the LLM-derived representations of the crystal structure description. Importantly, LLM can offer explanations and reasoning of the synthesizability for individual target structures, which are otherwise difficult in usual graph-based synthesizability predictions. The resulting models and explanations are demonstrated to better predict experimental outcomes than recent thermochemical predictions for synthesizability.

Results and Discussion

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[25,51] and began with the Materials Project (MP)[56] 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,[57] 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.[51] Detailed descriptions of the model, prompt, and fine-tuning process are in the Supporting Information. (Results obtained by fine-tuning the previous GPT-3.5 base model are 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,[51] 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 described graph-based crystal representation,[25] retrained on 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,[58] 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.[59,60] We adopted the same method for model evaluation and comparison in all cases.

As shown in Figure 1a, 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,[51] and the results here demonstrate the added value of including structural information. Several recent preprints[52,55,61] 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 Dec 2024, the cost to compute the text-embeddings is $0.065/M input 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.[58,62] 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 S10), 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[63,64] or on structural encodings,[65] 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.[66] A full exploration of this is outside the scope of the current article.

To investigate the two different (graph-based vs. GPT-embedding-based) structural representation capturability, we further evaluated both the PU-GPT-embedding and PU-CGCNN models across different description length divisions (<1000, 2000, 3000, …, 10000) of the hold-out-test dataset, as shown in Figure 1b. The result demonstrated that the model performances for each model 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. However, a recent preprint by Rubungo et al. came to the opposite conclusion—that LLM-based models excel with shorter textual descriptions, while CGCNN performs better on datasets with longer descriptions—for materials property prediction.[55] 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.

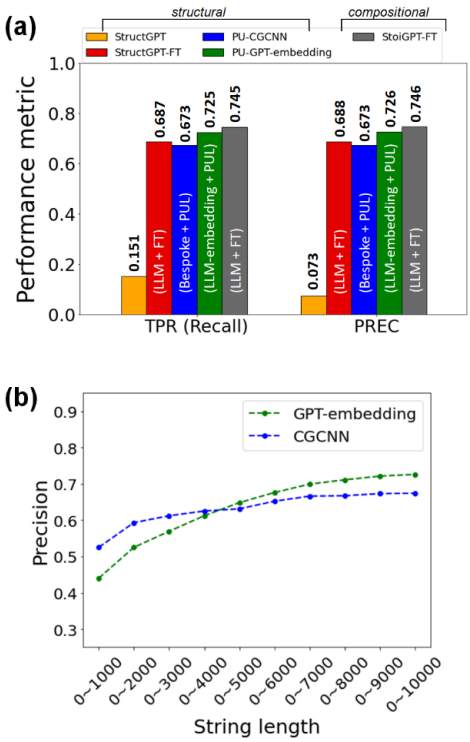


Figure 1. (a) Comparison of model performances for the general synthesizability prediction. FT indicates fine tuning and PUL indicates positive-unlabeled learning. (All calculated metrics are tabulated in Table S1.) (b) Model performances of PU-GPT-embedding and PU-CGCNN depending on the description length divisions of the hold-out-test dataset.

Structural sensitivity

To examine the sensitivity to input structure changes, we randomly varied the fractional coordinates within 1% and 5% in the CIF structures of the hold-out test set. These mutated CIF structures were then processed through Robocrystallographer to convert them into text descriptions.

As shown in Figure 2a, the overall text length increased, indicating that the structural symmetry was reduced during the mutation process, resulting in longer descriptions. We then investigated how the synthesizability prediction changes for these mutated structures. As shown in Figures 2b, the original structures exhibited an 71.0% recall for StructGPT-FT, but the 1% and 5% mutated structures showed a significantly lower recall of 3.1% and 1.2%, respectively. In the same context, for unlabeled data, the proportion predicted to be synthesizable dropped from 6.2% to 0.2% and 0.1%. PU-GPT-embedding model also exhibited a similar trend for the mutation test (Figure 2c), showing consistency with the StructGPT-FT case. This indicates that StructGPT-FT and PU-GPT-embedding models exhibit high sensitivity to even small structural changes for synthesizability prediction.

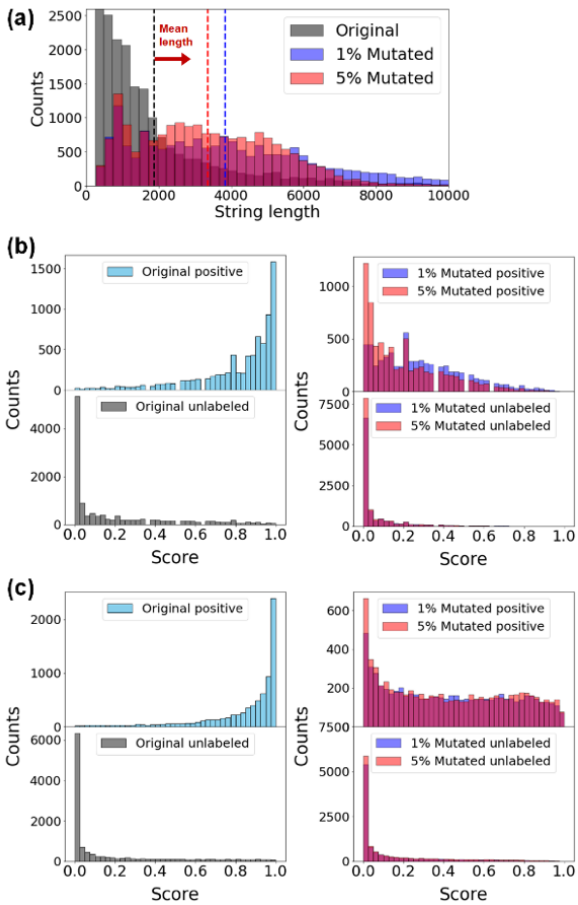


Figure 2. (a) The distribution of structural description length for the original and mutated crystal structures. (b) and (c) The result of synthesizability distribution by (b) StructGPT-FT and (c) PU-GPT-embedding model for original and 1%/ 5% mutated structures.

Comparison to thermodynamic stability

As an alternative to the purely data-driven approach used here, thermodynamics-based predictions assume that a material is synthesizable if its formation energy is within some threshold of the convex hull.[18,19,21-23] To compare the synthesizability through the thermodynamic stability, we obtained the energy above hull (Ehull) for each inorganic crystal from the Materials Project (MP); Figure S13 shows a summary histogram. Since Ehull < 0.05 eV/atom is often used as energetically favorable (near-stable) criterion, it has been considered as a crude estimate for crystal synthesizability.[24,70] Figure 3 shows a two-dimensional histogram of Ehull versus synthesizability scores of StructGPT-FT model. As shown in Figure 3a, the thermodynamic energy-based prediction achieved a recall of 87.1% in the metastable range (<0.2 eV/atom) and 74.4% in the near-stable range (<0.05 eV/atom) for the 7,491 hold-out positive compounds, which is comparable to StructGPT-FT score-based prediction (70.7%). However, the energy-based prediction showed that 33.3% and 72.0% of the hold-out unlabeled compounds are synthesizable by Ehull < 0.05 and 0.2 eV/atom criteria, respectively, which is significantly different from the result of the StructGPT-FT model, which shows only 6.1% (Figure 3b). This suggests that while thermodynamic approaches have good recall, they have much lower precision than our data driven approach, and will generate many false positives.

Recently, thermodynamic-based synthesizability predictions for twelve novel hypothetical compounds were tested in the laboratory.[24,70] Even though the these 12 compounds were all energetically near the ground state (<0.01 eV/atom), and therefore thermodynamically predicted to be “synthesizable”, repeated attempts to synthesize any of them failed. (These compounds are indicated as the stars in Figure 3b; additional details in Table S9.) In contrast, our StructGPT-FT model correctly assigns all of these compounds as negative, with synthesizability scores below the threshold (<0.777), in perfect agreement with the experimental outcomes. PU-GPT-embedding model also predicts their scores lower than the threshold (<0.813) for 10 out of 12 cases. (See the details in Table S9.) This indicates the strength of our LLM-based approach in capturing other aspects of metastability and synthetic accessibility that are neglected by the thermodynamic approach to synthesizability.

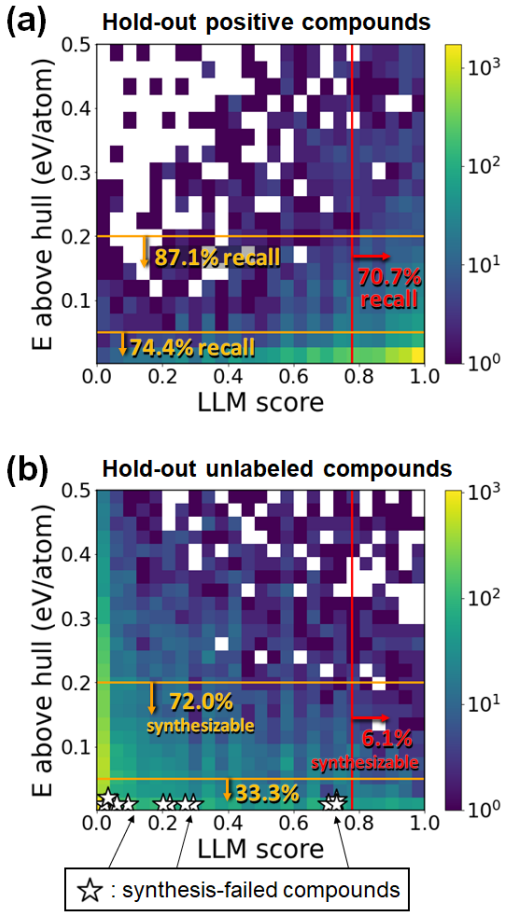


Figure 3. The two-dimensional colored histogram of energy above hull (Ehull) vs. score of StructGPT-FT model for (a) 7,491 hold-out test positive and (b) 12,204 hold-out test unlabeled materials. Red line indicates synthesizability threshold (0.777) and orange lines indicate thermodynamic stability thresholds for near-stable (<0.05 eV/atom) and metastable (<0.2 eV/atom) region. The white stars indicate the twelve hypothetical compounds, which were demonstrated as non-synthesizable by several attempted experiments. (See the details for these materials in Table S9.)

Explanation / Inference for synthesizability

Using the synthesizability predictions made above, we then used the StructGPT-FT 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 StructGPT-FT with a total of 17,734 structure-prediction pairs, where the predictions of two models (StructGPT-FT and 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. StructGPT-FT usually answered the explana

tion 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 StructGPT-FT use in these generated reasons? We accumulated all the keywords of reasons for the 17,734 structures, and put top-500 frequently mentioned raw keywords into GPT-4o to cluster similar keywords and make 10 representative categories. (See the detailed keyword items in the Supporting Information.) We plotted a bar histogram of the 10 most relevant reasons (Figure 4a). It is generally acknowledged that thermodynamic stability alone is an insufficient factor for material synthesizability. In our results, explanations for about 25.4% (4,513/17,734) of structures included thermodynamic stability as a reason for their synthesizability. To investigate how the importance of each factor changes according to the structure type, we analyzed the explanation factor proportion of top-3 frequent structure types; cubic perovskites, Heusler compounds, and spinel structures. Figure 4b showed the structural t-SNE distribution, with top-10 structure types highlighted by colors. The top-3 structure types exhibited different proportions of explanation keywords, suggesting the importance of each factor varies by substance type (e.g., bonding and coordination characteristics for cubic perovskites, inconsistencies and chemical compatibility for Heusler compounds, and space group and symmetry for spinel structures). The result for the remaining top-10 structure types were shown in Figure S7 in the Supporting Information. Figure 4c showed the structural t-SNE distribution colored by their synthesizability. When considering the distribution based on structure types in Figure 4b, it becomes evident that the synthesizability may vary depending on the structure types (e.g., most Heusler compounds and spinel-type structures are predicted non-synthesizable).

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 and S7. 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.[67] Interestingly, while geometry and bond-length are the most commonly invoked reasons by the model, they actually have less impact on the final prediction.

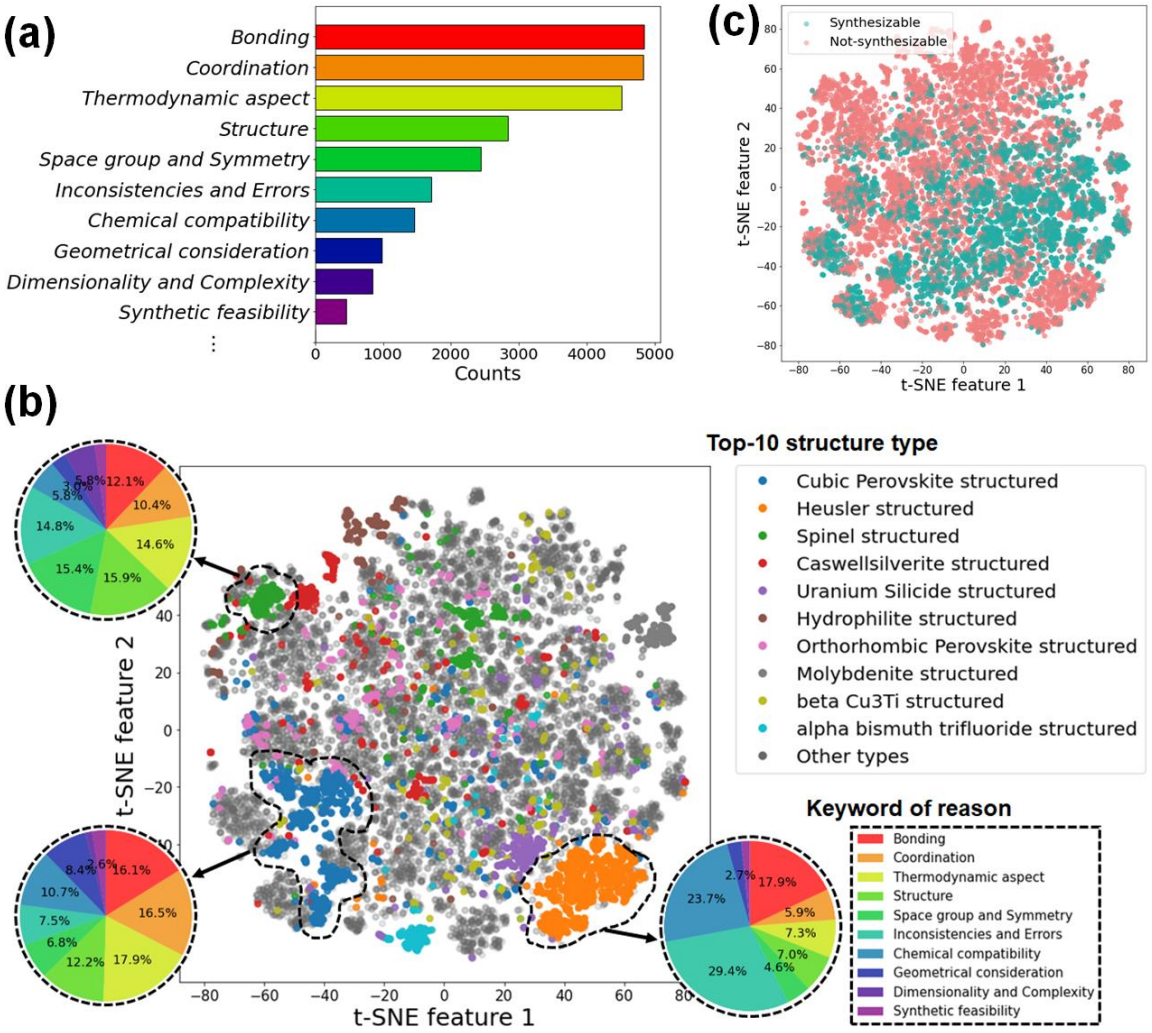


Figure 4. (a) 10 most relevant reasons for general synthesizability. (b) t-SNE for crystal structures in which top-10 structure types were colored to indicate their distribution. Top-3 types (blue for cubic perovskites, orange for heusler compounds, and green for spinel structures) showed their explanation proportion by pie-charts. (c) t-SNE for crystal structures colored by their synthesizability.

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 S

c(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). 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.[68] Finally, by the classical logical Principle of Non-Contradiction,[69] 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 5a, 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 5b) and the arithmetic mean (Figure 5c) 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 S12e), 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 StructGPT-FT use principles that are generally well-attested by the training corpus and are internally consistent.

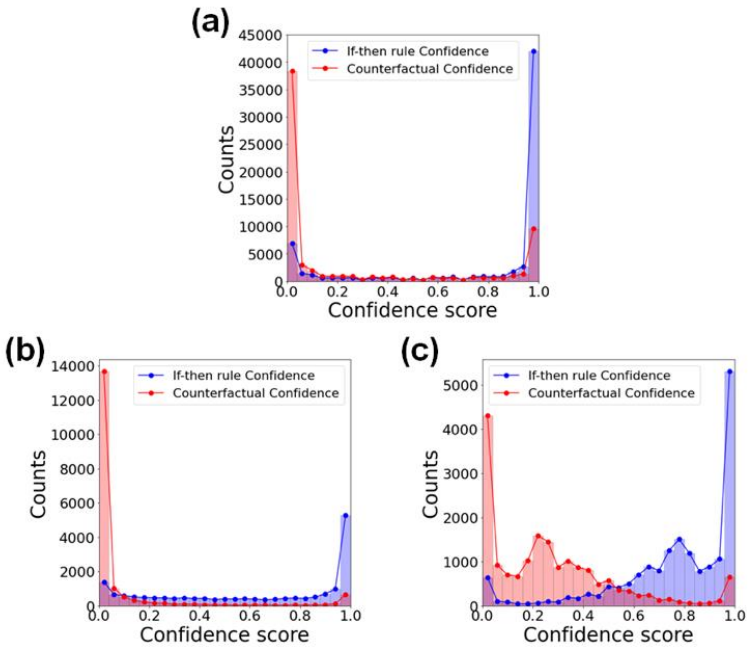


Figure 5. The results of assessing the whole explanations based on the log-probability. Since StructGPT-FT usually answered the explanation with 4 or 5 reasons per 1 material (Figure S6), we combined (a) each explanation probability by (b) geometric mean of probabilities and (c) arithmetic mean of probabilities.

Transfer learning to perovskites

Transfer learning is a machine learning technique where a model developed for a particular task is reused as the starting point for a model on a second related task, effectively leveraging pre-existing knowledge to improve learning efficiency and performance. To demonstrate the feasibility of transfer-learning for the LLM-based method presented above, we considered the problem of perovskite synthesizability, using a separate dataset of 1,533 synthesized and 13,276 hypothetical perovskite structures, Full methods and results are discussed in Supporting Information. The overall model quality results for this transfer learning setting were comparable to the more general case discussed above, although the CGCNN method had the best performance. This is consistent with the performance trend seen in Figure 1b, as the perovskites have simpler structures and more concise text descriptions (Figure S2b), and CGCNN tends to do better in this regime.

Conclusion

We utilized LLMs for structure-based general 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,[71] 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.

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.

Through comparisons with thermodynamic energy-based predictions using both statistical analysis and specific examples, it was observed that relying solely on thermodynamic-based predictions can result in many false positive cases. Since many factors influence the synthesizability of materials, LLM-based predictions and explanations can be helpful for understanding the complex material chemistry.

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

(1) Since this model highly relies on existing material database, the prediction could be biased by the distribution of already-synthesized materials.[72-74] 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.[75] Furthermore, there might be errors in the material database, making it necessary to carefully validate the data obtained from DFT calculations, which we used as training data for our data-driven method.

(2) In this study, we focused on ordered inorganic crystal structures. Defects and disorder are inevitable in real world materials,[76] indicating the need for future research that can address these aspects as well.

(3) We used general purpose 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.[77,78] Alternatively, introducing material-specific latent vectors through unsupervised learning of text taken from abstracts

in the materials science literature,[79,80] Robocrystallographer structure descriptions,[81,82] or directly on the text of CIF files[83] 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, RS-2024-00464386) 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 https://zenodo.org/records/14729225. 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

References

[1] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, Nature 2017, 543, 657-664.

[2] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, Nature Chemistry 2009, 1, 37-46.

[3] A. Zunger, Nature Reviews Chemistry 2018, 2, 0121.

[4] M. Jansen, Advanced Materials 2015, 27, 3229-3242.

[5] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, Nature Materials 2013, 12, 191-201.

[6] S. Muy, J. Voss, R. Schlem, R. Koerver, S. Sedlmaier, F. Maglia, P. Lamp, W. Zeier, Y. Shao-Horn, iScience 2019, 16, 270-282.

[7] Y. Zhuo, A. Mansouri Tehrani, A. O. Oliynyk, A. C. Duke, J. Brgoch, Nature Communications 2018, 9, 4377.

[8] J. Zhou, L. Shen, M. Yang, H. Cheng, W. Kong, Y. P. Feng, Chemistry of Materials 2019, 31, 1860-1868.

[9] T. Mueller, G. Hautier, A. Jain, G. Ceder, Chemistry of Materials 2011, 23, 3854-3862.

[10] D. H. Mok, S. Back, arXiv preprint 2024, DOI: 10.48550/arXiv.2407.14040.

[11] J. Yano, K. J. Gaffney, J. Gregoire, L. Hung, A. Ourmazd, J. Schrier, J. A. Sethian, F. M. Toma, Nature Reviews Chemistry 2022, 6, 357-370.

[12] S. Back, A. Aspuru-Guzik, M. Ceriotti, G. Gryn'ova, B. Grzybowski, G. H. Gu, J. Hein, K. Hippalgaonkar, R. Hormázabal, Y. Jung, Digital Discovery 2024, 3, 23-33.

[13] A. K. Singh, J. H. Montoya, J. M. Gregoire, K. A. Persson, Nature communications 2019, 10, 443.

[14] J. Noh, D.-W. Jeong, K. Kim, S. Han, M. Lee, H. Lee, Y. Jung, International Conference on Machine Learning 2022, 16952-16968.

[15] S. Chen, Y. Jung, Journal of Cheminformatics 2024, 16, 83.

[16] H. Park, A. Onwuli, K. Butler, A. Walsh, Faraday Discussions 2024, DOI: 10.1039/D4FD00063C.

[17] N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, Z. Ulissi, arXiv preprint 2024, DOI: 10.48550/arXiv.2402.04379.

[18] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, G. Ceder, Science advances 2016, 2, e1600225.

[19] M. Aykol, S. S. Dwaraknath, W. Sun, K. A. Persson, Science advances 2018, 4, eaaq0148.

[20] M. Aykol, J. H. Montoya, J. Hummelshøj, Journal of the American Chemical Society 2021, 143, 9244-9259.

[21] W. Ye, C. Chen, Z. Wang, I.-H. Chu, S. P. Ong, Nature Communications 2018, 9, 3800.

[22] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, G. Ceder, npj Computational Materials 2020, 6, 97.

[23] C. J. Bartel, A. W. Weimer, S. Lany, C. B. Musgrave, A. M. Holder, npj Computational Materials 2019, 5, 4.

[24] J. H. Montoya, C. Grimley, M. Aykol, C. Ophus, H. Sternlicht, B. H. Savitzky, A. M. Minor, S. B. Torrisi, J. Goedjen, C. C. Chung, A. H. Comstock, S. Sun, Chemical Science 2024, 15, 5660-5673.

[25] J. Jang, G. H. Gu, J. Noh, J. Kim, Y. Jung, Journal of the American Chemical Society 2020, 142, 18836-18843.

[26] N. C. Frey, J. Wang, G. I. n. Vega Bellido, B. Anasori, Y. Gogotsi, V. B. Shenoy, ACS Nano 2019, 13, 3031-3041.

[27] J. Jang, J. Noh, L. Zhou, G. H. Gu, J. M. Gregoire, Y. Jung, Matter 2024, 7, 2294-2312.

[28] E. R. Antoniuk, G. Cheon, G. Wang, D. Bernstein, W. Cai, E. J. Reed, npj Computational Materials 2023, 9, 155.

[29] R. Zhu, S. I. P. Tian, Z. Ren, J. Li, T. Buonassisi, K. Hippalgaonkar, ACS Omega 2023, 8, 8210-8218.

[30] A. Davariashtiyani, Z. Kadkhodaie, S. Kadkhodaei, Communications Materials 2021, 2, 115.

[31] F. Mordelet, J.-P. Vert, Pattern Recognition Letters 2014, 37, 201-209.

[32] J. Bekker, J. Davis, Machine Learning 2020, 109, 719-760.

[33] G. H. Gu, J. Jang, J. Noh, A. Walsh, Y. Jung, npj Computational Materials 2022, 8, 71.

[34] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Nature 2018, 559, 547-555.

[35] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Science Robotics 2019, 4, eaay7120.

[36] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Pattern Recognition 2017, 65, 211-222.

[37] S. Yu, N. Ran, J. Liu, Artificial Intelligence Chemistry, 2024, 2(2), 100076.

[38] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, P. Schwaller, Nature Machine Intelligence 2024, 1-11.

[39] D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Nature 2023, 624, 570-578.

[40] M. C. Ramos, C. J. Collison, A. D. White, arXiv preprint 2024, DOI: 10.48550/arXiv.2407.01603.

[41] G. Lei, R. Docherty, S. J. Cooper, Digital Discovery 2024, 3, 1257-1272.

[42] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, Journal of the American Chemical Society 2023, 145, 18048-18062.

[43] J. Schrier, Journal of Chemical Education 2024, 101, 1782-1784.

[44] A. Mirza, N. Alampara, S. Kunchapu, B. Emoekabu, A. Krishnan, M. Wilhelmi, M. Okereke, J. Eberhardt, A. M. Elahi, M. Greiner, arXiv preprint 2024, DOI: 10.48550/arXiv.2404.01475.

[45] H. Wang, K. Li, S. Ramsay, Y. Fehlis, E. Kim, J. Hattrick-Simpers, arXiv preprint 2024, DOI: 10.48550/arXiv.2409.14572.

[46] C, Bajan, G. Lambard, Digital Discovery 2024, accepted, DOI: 10.1039/D4DD00319E.

[47] T. Dinh, Y. Zeng, R. Zhang, Z. Lin, M. Gira, S. Rajput, J.-y. Sohn, D. Papailiopoulos, K. Lee, Advances in Neural Information Processing Systems 2022, 35, 11763-11784.

[48] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, Nature Machine Intelligence 2024, 6, 161-169.

[49] Z. Xie, X. Evangelopoulos, Ö. H. Omar, A. Troisi, A. I. Cooper, L. Chen, Chemical Science 2024, 15, 500-510.

[50] S. Zhong, X. Guan, Environmental Science & Technology Letters 2023, 10, 872-877.

[51] S. Kim, Y. Jung, J. Schrier, Journal of the American Chemical Society 2024, 146, 19654-19659.

[52] Z. Song, S. Lu, M. Ju, Q. Zhou, J. Wang, arXiv preprint 2024, DOI: 10.48550/arXiv.2407.07016.

[53] R. Jacobs, M. P. Polak, L. E. Schultz, H. Mahdavi, V. Honavar, D. Morgan, arXiv preprint 2024, DOI: 10.48550/arXiv.2409.06080.

[54] J. V. Herck, M. V. Gil, K. M. Jablonka, A. Abrudan, A. Anker, M. Asgari, B. Blaiszik, A. Buffo, L. Choudhury, C. Corminboeuf, H. Daglar, A. M. Elahi , I. T. Foster , S. García , M. Garvin , G. Godin , L. L. Good , J. Gu , N. X. Hu , X. Jin , T. Junkers , S. Keskin , T. Knowles , R. Laplaza , M. Lessona , S. Majumdar , H. Mashhadimoslem , R. D McIntosh , S. M. Moosavi , B. Mouriño , F. Nerli , C. Pevida , N. Poudineh , M. R. Kochi , K. Saar , F. H. Saboor , M. Sagharichiha , K. J. Schmidt , J. Shi , E. Simone , D. Svatunek , M. Taddei , I. V. Tetko, D. Tolnai , S. Vahdatifar , J. K. Whitmer , F. Wieland , R. W. Römer , A. Züttel, B. Smit, Chemical Science 2024, accepted, DOI: 10.1039/D4SC04401K.

[55] A. N. Rubungo, K. Li, J. Hattrick, A. B. Dieng, arXiv preprint 2024, DOI: 10.48550/arXiv.2411.00177.

[56] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, APL Materials 2013, 1, 011002.

[57] A. M. Ganose, A. Jain, MRS Communications 2019, 9, 874-881.

[58] OpenAI, “New embedding models and API updates”, can be found under https://openai.com/index/new-embedding-models-and-api-updates, 2024, (accessed: Dec. 2, 2024).

[59] S. Jain, M. White, P. Radivojac, Proceedings of the AAAI Conference on Artificial Intelligence 2017, 31.

[60] D. Zeiberg, S. Jain, P. Radivojac, Proceedings of the AAAI Conference on Artificial Intelligence 2020, 34, 6729-6736.

[61] N. Alampara, S. Miret, K. M. Jablonka, arXiv preprint 2024, DOI: 10.48550/arXiv.2406.17295.

[62] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, Advances in Neural Information Processing Systems 2022, 35, 30233-30249.

[63] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, Chemistry of Materials 2015, 27, 735-743.

[64] M. Kuban, Š. Gabaj, W. Aggoune, C. Vona, S. Rigamonti, C. Draxl, MRS Bulletin 2022, 47, 991-999.

[65] S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, F. Pan, Wiley Interdisciplinary Reviews: Computational Molecular Science 2022, 12, e1558.

[66] D. H. Rouvray, Molecular Similarity I 2005, 1-30.

[67] A. N. Rubungo, C. Arnold, B. P. Rand, A. B. Dieng, arXiv preprint 2023, DOI: 10.48550/arXiv.2310.14029.

[68] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, arXiv preprint 2022, DOI: 10.48550/arXiv.2207.05221.

[69] L. R. Horn, "Contradiction", The Stanford Encyclopedia of Philosophy (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2024/entries/contradiction/>, (accessed: Dec. 2, 2024)

[70] W. Ye, X. Lei, M. Aykol, J. H. Montoya, Scientific Data 2022, 9, 302.

[71] Q. Liu, M. P. Polak, S. Y. Kim, M. Shuvo, H. S. Deodhar, J. Han, D. Morgan, H. Oh, arXiv preprint 2024, DOI: 10.48550/arXiv.2409.06756.

[72] W. Sun, N. David, Faraday Discussions 2024, DOI: 10.1039/D4FD00112E.

[73] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, Nature 2019, 573, 251-255.

[74] J. Schrier, A. J. Norquist, T. Buonassisi, J. Brgoch, Journal of the American Chemical Society 2023, 145, 21699-21716.

[75] K. Li, A. N. Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A. B. Dieng, J. Hattrick-Simpers, arXiv preprint 2024, DOI: 10.48550/arXiv.2406.06489.

[76] A. K. Cheetham, R. Seshadri, Chemistry of Materials 2024, 36, 3490-3495.

[77] T. Aarsen, "Training and Finetuning Embedding Models with Sentence Transformers v3", can be found under https://huggingface.co/blog/train-sentence-transformers, 2024, (accessed: Dec. 2, 2024).

[78] P. Schmid, "Fine-tune Embedding models for Retrieval Augmented Generation (RAG)", can be found under https://www.philschmid.de/fine-tune-embedding-model-for-rag, 2024, (accessed: Dec. 2, 2024).

[79] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Nature 2019, 571, 95-98.

[80] B. Zhang, H. Xiao, G. Ye, Z. Song, T. Han, E. Sharman, M. Luo, A. Cheng, Q. Zhu, H. Zhao, The Journal of Physical Chemistry Letters 2023, 15, 212-219.

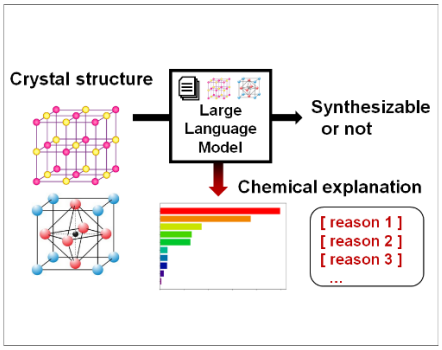
[81] H. M. Sayeed, S. G. Baird, T. D. Sparks, ChemRxiv preprint 2023, DOI: 10.26434/chemrxiv-2023-3q8wj.

[82] J. Qu, Y. R. Xie, K. M. Ciesielski, C. E. Porter, E. S. Toberer, E. Ertekin, npj Computational Materials 2024, 10, 58.

[83] L. Yadav, AIP Advances 2024, 14, 045205.

[84] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Computational Materials 2015, 1, 15010.

Entry for the Table of Contents



We utilized LLMs for structure-based general synthesizability prediction along with their explanations. Fine-tuned LLMs and LLM-embedding-based bespoke ML models showed promising performance compared to the traditional graph-based bespoke ML models. Furthermore, fine-tuned LLMs can provide explainability by inferring the reasons for determining the synthesizability with simple prompts.