Volume 4
Number 6
June 2025
Pages 1375-1652

Digital Discovery

rsc.li/digitaldiscovery

ISSN 2635-098X

ROYAL SOCIETY
OF CHEMISTRY

PAPER
Philipp Benner *et al.*
SynCoTrain: a dual classifier PU-learning framework for synthesizability prediction

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 1437

SynCoTrain: a dual classifier PU-learning framework for synthesizability prediction†

Sasan Amariamir,[a] Janine George[ab] and Philipp Benner*[a]

Received 14th December 2024
Accepted 23rd March 2025

DOI: 10.1039/d4dd00394b

rsc.li/digitaldiscovery

Material discovery is a cornerstone of modern science, driving advancements in diverse disciplines from biomedical technology to climate solutions. Predicting synthesizability, a critical factor in realizing novel materials, remains a complex challenge due to the limitations of traditional heuristics and thermodynamic proxies. While stability metrics such as formation energy offer partial insights, they fail to account for kinetic factors and technological constraints that influence synthesis outcomes. These challenges are further compounded by the scarcity of negative data, as failed synthesis attempts are often unpublished or context-specific. We present SynCoTrain, a semi-supervised machine learning model designed to predict the synthesizability of materials. SynCoTrain employs a co-training framework leveraging two complementary graph convolutional neural networks: SchNet and ALIGNN. By iteratively exchanging predictions between classifiers, SynCoTrain mitigates model bias and enhances generalizability. Our approach uses Positive and Unlabeled (PU) learning to address the absence of explicit negative data, iteratively refining predictions through collaborative learning. The model demonstrates robust performance, achieving high recall on internal and leave-out test sets. By focusing on oxide crystals, a well-characterized material family with extensive experimental data, we establish SynCoTrain as a reliable tool for predicting synthesizability while balancing dataset variability and computational efficiency. This work highlights the potential of co-training to advance high-throughput materials discovery and generative research, offering a scalable solution to the challenge of synthesizability prediction.

1 Introduction

Material discovery is a foundational pillar of modern science and perhaps the driving motivation behind materials science. It supports advancements in numerous scientific and technological disciplines. In this field, the ability to predict synthesizability is crucial. Developing materials with novel properties expands the possibilities in endeavors from functional materials used in biomedical devices to addressing the challenges of climate change.[1] In the past decade or so, efforts such as the Materials Genome Initiative aimed to accelerate the discovery, development, and deployment of new materials in the hopes of societal betterment.[1,2] An essential part of realizing this goal is employing high-throughput simulations and experiments for screening candidate materials with desirable properties.[1,3] Unfortunately, a substantial amount of resources and effort can be wasted on hypothetical materials that currently cannot be synthesized.

Historically, physico-chemical based heuristics such as the Pauling Rules[4] or the charge-balancing criteria[5] have been used to assess materials stability and synthesizability. Nevertheless, these simplified approaches have been shown to be insufficient, as more than half of the experimental (already synthesized) materials on the Materials Project database[6] do not meet these criteria for synthesizability.[5,7]

In more recent attempts, material scientists often employed thermodynamic stability as a proxy for synthesizability, ignoring the effect of kinetic stabilization. This involves conducting first-principle calculations to estimate the formation energy of crystals and their distance from the convex hull. A negative formation energy, or a minimal distance from the convex hull, is commonly interpreted as an indicator of synthesizability.[8–12] While stability significantly contributes to synthesizability, it is just one aspect of this complex issue. There are many –potentially interesting– metastable materials that do exist, even though their formation energies deviate from the ground-state.[8,11,13–15] These materials can be synthesized in alternate thermodynamic conditions in which they are the ground-state. After removing the favorable thermodynamic field, they have stayed stuck in the metastable structure by kinetic stabilization.[8] On the other hand, there are many hypothetical stable materials in well-explored chemical spaces

[a] *Federal Institute of Materials Research and Testing, Unter den Eichen 87, 12205 Berlin, Germany. E-mail: philipp.benner@bam.de*

[b] *Friedrich Schiller University Jena, Institute of Condensed Matter Theory and Solid-State Optics, Max-Wien-Platz 1, 07743 Jena, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00394b

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

which have never been synthesized. This could be due to a high activation energy barrier between them and the common precursors.[13–15] Beyond the theoretical and thermodynamic considerations, synthesizability is also a technological problem. Novel materials that are developed through cutting-edge methods were practically unsynthesizable before the invention of their methods of synthesis. For example, new high-entropy alloys with great potential for catalysis applications were recently synthesized using the Carbothermal Shock (CTS) method.[16] Their particular homogeneous components and uniform structures were not accessible through conventional synthesis methods. On the other hand, some materials can only be synthesized under specific conditions, such as extremely high pressures.[17]

The fact that estimating synthesizability is related to materials structures without a straightforward formula to solve for, makes it an apt candidate for machine learning. This and many other challenges have made machine learning the ideal technique to accelerate material discovery.[18] In this work, we define a classification task for two classes of materials, namely synthesizable (the positive class), and unsynthesizable (the negative class). This classification comes with a few challenges and intricacies. The first one is encoding materials structures in a machine understandable format. Some previous works have circumvented this challenge in creative ways such as combining different elemental features,[15,19] using text-mining algorithms to search the relevant literature to identify synthesizable materials,[20] using the picture of crystal cells with convolutional neural networks,[21] or even a network analysis of materials discovery timeline with respect to their stability.[22] Others,[14,23] including this work, utilize graph convolutional neural networks (GCNNs) to encode and learn from crystal structures. While the GCNNs are more complicated to implement, they have the advantage of including more information about the structure than composition alone or the other previously mentioned approaches that represent the structure information indirectly through a proxy.

The second challenge of estimating synthesizability lies within the nature of the available data. Unlike a typical classification task, we do not have access to enough negative data. On the one hand, this is due to the fact that unsuccessful attempts of synthesis are not typically published nor uploaded to public databases. The attempts of using such failed experiments[24] inevitably remain confined to local labs and a small class of materials. Also, synthesis success strongly depends on the synthesis conditions and technology. Hence, the failure of synthesis attempts in one setting does not necessarily imply failure in a different lab with different synthesis methods or equipment. Finally, creating a proper negative-set for training a classifier is a whole new challenge.[5] If the negative-set is too different from real materials, it may not teach the model a meaningful decision boundary for detecting synthesizability. To design a realistic-looking negative-set, one would need to understand the features that determine synthesizability in the first place.

The final challenge in this task comes as a fundamental aspect of machine learning. Regardless of which model is chosen, it will inherently exhibit a certain degree of bias. One introduces a possibly unintended bias when selecting one model over another, since the model's ability to generalize out of sample is, in part, predetermined by its architecture. This model bias comes even with the best performing models. In fact, a model with great benchmarks might perform worse than simpler models when predicting targets for out-of-distribution data,[25] perhaps due to overfitting. This challenge becomes particularly pronounced when predicting synthesizability. The objective is to forecast a target for new and often out-of-distribution data, where the issue of generalization is most acute. The lack of the negative data compounds this issue, as it makes performance metrics less reliable. One way to mitigate this issue is by leveraging multiple models. An ensemble of models with diverse architectures and learning strategies can help balance individual model biases, improve robustness, and provide a more reliable assessment of synthesizability. By aggregating predictions from multiple models, the approach reduces overfitting, enhances generalization, and compensates for the missing negative data, leading to more accurate and trustworthy synthesizability predictions.[25]

To address these challenges, we have developed a model ready for integration into high-throughput simulations and generative materials research. It is called SynCoTrain (pronounced similar to 'Synchrotron'). It is a semi-supervised classification model designed for predicting synthesizability of oxide crystals. SynCoTrain addresses the generalizability issue by utilizing co-training. Co-training is an iterative semi-supervised learning process designed for scenarios with some positive data and a lot of unlabeled data.[26,27] It leverages the predictive power of two distinct classifiers to find and label positive data points among the unlabeled data. Different models have different biases, and by combining their predictions, we can practically reduce these biases while keeping what they learn about the target. We use the Atomistic Line Graph Neural Network (ALIGNN)[28] and the SchNetPack[29,30] models as our chosen classifiers. They are both innovative GCNNs with distinct attributes. ALIGNN is unique in that it directly encodes both atomic bonds and bond angles into its architecture, offering a perspective that aligns with a chemist's view of the data. SchNetPack stands out for using a unique continuous convolution filter which is suitable for encoding atomic structures, which can be thought of as a physicist's perspective on the data.

At each step of co-training, SynCoTrain learns the distribution of the synthesizable crystals through the Positive and Unlabeled learning (PU learning) method introduced by Mordelet and Vert.[31] This base PU learning method with a different classifier has already been employed to predict synthesizability for all classes of crystals[14] and for perovskites specifically.[23] In this work, we utilize multiple PU learners as the building blocks for co-training. In each iteration of co-training, the learning agents exchange the knowledge they gained from the data between each other. Eventually, the labels are decided based on average of their predictions. This process increases the prediction reliability and accuracy, much like two experts who discuss and reconcile their views before finalizing a complex decision.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

This collaborative approach suggests that co-training is more likely to generalize effectively to unseen data compared to using a single model with equivalent classification metrics such as accuracy or recall.

We verify the performance of the model by recall for an internal test-set and a leave-out test-set. We also evaluate our model further by predicting whether a crystal is stable or not for the same data points. Note that in predicting stability, we do not aim for a good performance. In fact, we expect an overall poor performance due to high contamination of the unlabeled data;[31] more info in ESI.† However, we compare the ground truth recall in stability to the recall produced by the PU learning, to gauge the reliability of the latter.

We chose a single family of materials, oxides, to establish the utility of co-training in predicting materials properties. Oxides are a well-studied class of materials with a large amount of experimental data to learn from ref. 32 and 33. A higher number of training data would typically decrease the classification error in machine learning. However, training across all available families of crystals would introduce greater variability in the dataset, potentially increasing the uncertainty and error margins in our results. In other words, the prediction quality for new materials would vary substantially. By achieving high recall values with oxides as our training data, we demonstrate the effectiveness of co-training. This approach ensures reliable results while maintaining reasonable training times for our models. Our data stems from the Materials Project database,[6] in which all of the crystal structures have been optimized with DFT and should be of similar quality. In many cases, the starting structures for optimization were those from the Inorganic Crystal Structure Database (ICSD).[34] For training machine learning models, it is crucial to minimize obvious biases, which can arise from combining data from different sources. Such biases can be easily detected by machine learning models, leading to distorted performance metrics.[35] To mitigate this risk, we rely exclusively on a single data source for training our model.

2 Results and discussion

2.1 Model development

The data for oxide crystals were obtained through the Materials Project API. The experimental and theoretical data are distinguished based on the 'theoretical' attribute. We used the get_valences function of pymatgen[36] to include only oxides where the oxidation number is determinable and the oxidation state of oxygen is −2.

Less than 1% of the experimental data with energy above hull higher than 1 eV were removed, as potentially corrupt data. The learning began with 10 206 experimental and 31 245 unlabeled data points.

Co-training consists of two separate iteration series, the results of which are averaged in the final step. In the first series, we start by training a base PU learner with an ALIGNN classifier. This is the iteration '0' of co-training, and this step is called ALIGNN0. The learning agent predicts positive labels for some of the unlabeled data, creating a pseudo-positive class. This class is added to the original experimental data, expanding the initial positive class. Iteration '1' of co-training on this series is to train a base PU learner with the other classifier, here the SchNet, on the newly expanded labels. This step is called coSchnet1. Each iteration provides newly expanded labels for the next iteration. The classifiers alternate for each iteration, from ALIGNN to SchNet and *vice versa*, as shown in Fig. 1a.

Parallel to this series, we set up a mirror series where iteration '0' begins with a SchNet based PU learner. This step of iteration '0' is called SchNet0. This series learns the data from a different, complementary view compared to the former series, see Fig. 1a. It continues in the same manner with alternating classifiers. The order of the steps in each series can be found in Table 1.

Each base PU learner produces a synthesizability score between 0 and 1 for each unlabeled datum. This is done through 60 runs of the bagging method established by Mordelet and Vert,[31] as illustrated in Fig. 1c. In each independent run of this ensemble learner, a random subset of the unlabeled data is sampled to play the role of the negative data in training the classifier. The average of the predictions in these runs for data points that were not part of the training in that run yields the synthesizability score. This score is interpreted as the predicted probability of being synthesizable. A threshold of 0.5 is applied for labeling each datum as either synthesizable (labeled 1) or not-synthesizable (labeled 0).

After several iterations of co-training, the optimal iteration is chosen based on the prediction metrics (*i.e.* recall rate). Continuing to further iterations yields diminishing returns in performance metric while risking reinforcing existing model bias. The scores provided from the two series at the optimal iteration are then averaged. The 0.5 cutoff threshold is applied to this averaged score to produce the final synthesizability score. Once we have synthesizability labels for both the experimental and theoretical data, a simple machine learning task remains. We train a classifier on these labels and end up with a model that can predict synthesizability (see Fig. 1b).

2.2 Model evaluation and results

Within each single run of the base PU learning, the classifiers optimize for accuracy, as they operate unaware of the PU nature of the data. When reporting the performance of PU learning, however, one should not use accuracy, Precision or the F1-score. These common measures assume knowledge of the negative labels and a false positive count. We use recall, also known as sensitivity or true positive rate (true positive/(true positive + false negative)) to report and benchmark the performance of our PU learner as it only relies on the knowledge of the positive data.

In our study, we employ two distinct test-sets to measure recall. The first is a dynamic test-set, which varies with each iteration of base PU learning. The second is a leave-out test-set that remains untouched during all training iterations. As the result, we obtain a 'recall range' between the two distinct recall measures; an averaged recall for the dynamic test-set and a leave-out recall. This gives us more information than a single

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

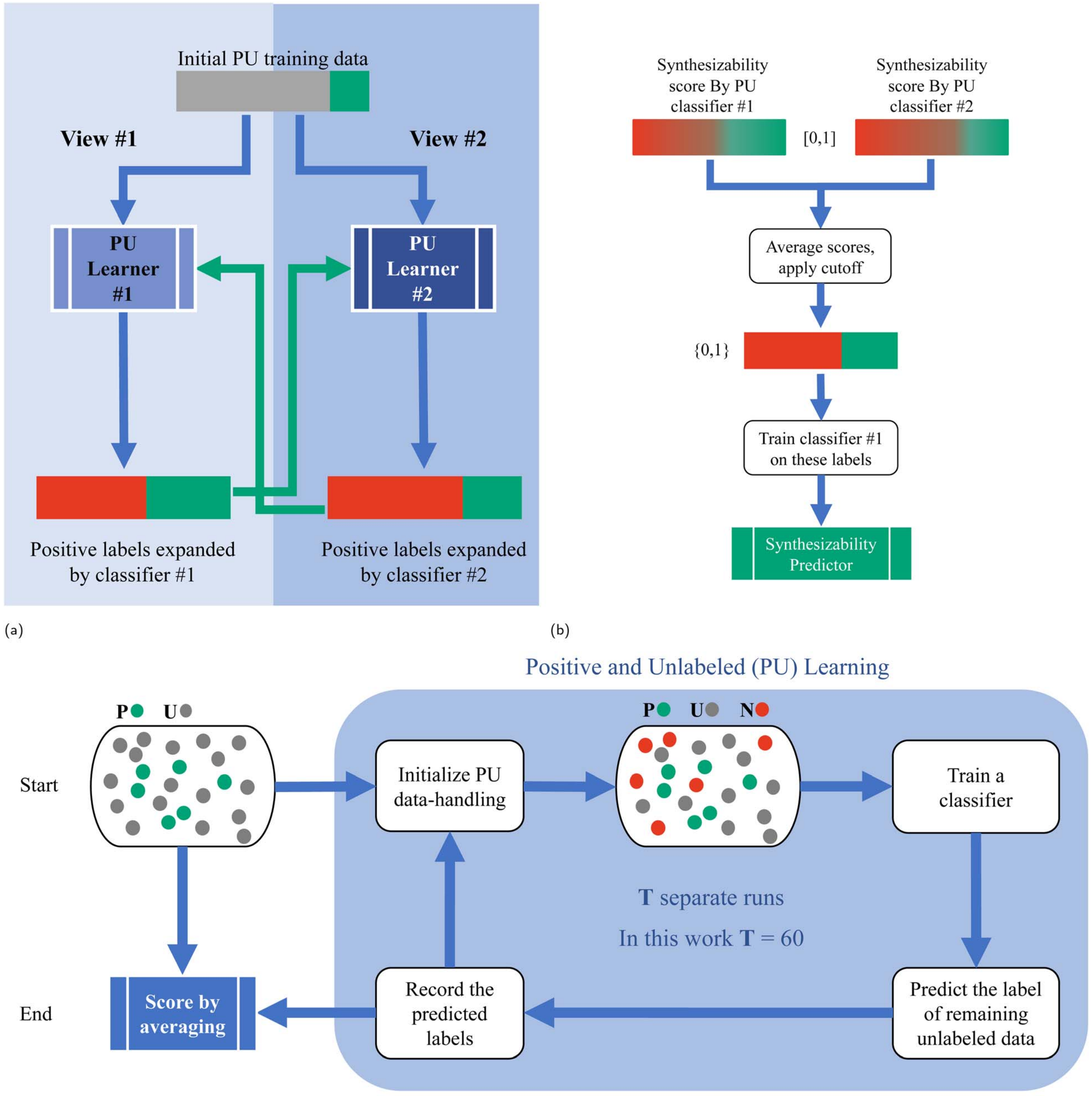

Initial PU training data
View #1
View #2
PU Learner #1
PU Learner #2
Positive labels expanded by classifier #1
Positive labels expanded by classifier #2
(a)
Synthesizability score By PU classifier #1
Synthesizability score By PU classifier #2
[0,1]
Average scores, apply cutoff
{0,1}
Train classifier #1 on these labels
Synthesizability Predictor
(b)
Positive and Unlabeled (PU) Learning
P U
P U N
Start
Initialize PU data-handling
Train a classifier
T separate runs
In this work T = 60
Predict the label of remaining unlabeled data
Record the predicted labels
Score by averaging
End
(c)

Fig. 1 Overview of the Workflow in SynCoTrain (a) the PU data is passed to two distinct PU classifiers, each learning from a different view of the data. Each classifier labels unlabeled data points as positive or negative. The new labels from each PU classifier are used to expand the positive class for retraining the other classifier. (b) After co-training steps, each unlabeled data point receives a prediction score from each PU classifier. An average of these scores is calculated for each data point, and a cutoff is applied to produce a label. All the data, now labeled, are used to train a final classifier to predict synthesizability. (c) The PU learning process. Positive, negative, and unlabeled data are depicted as green, red, and gray circles respectively. Each run starts with training a classifier, with a randomly chosen subset of the unlabeled data used as the negative class. Labels are predicted for the remaining unlabeled data, and final scores are computed by averaging these predictions.

recall value. The construction and reasoning behind this are detailed in the ground truth evaluation section.

The recall values for each iteration are depicted in Fig. 2. The two distinct co-training series are separately visualized to clearly illustrate recall changes at each step. Iteration ‘0’ represents a basic PU learning approach with isolated classifiers, without any co-training. We see that the SchNet0 series somewhat plateaus in iteration ‘2’, while the ALIGNN0 series still improves in recall. However, neither series make significant improvement on their recall in iteration ‘3’. This suggests that using the third iteration yields diminishing returns in terms of new learning, while risking enforcing models' biases through too many repetitions. Furthermore, the predicted positive rate increases in both series for iteration ‘3’, without a meaningful increase in

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 1 Co-training steps

Co-training steps	Iteration ‘0’		Iteration ‘1’		Iteration ‘2’		Iteration ‘3’	Averaging scores
Training data source	Original labels		Labels expanded by Iteration ‘0’		Labels expanded by Iteration ‘1’		Labels expanded by Iteration ‘2’	Scores provided by the optimal iteration
Training series	ALIGNN0	>	coSchnet1	>	coAlignn2	>	coSchnet3	Synthesizability scores
	SchNet0	>	coAlignn1	>	coSchnet2	>	coAlignn3	

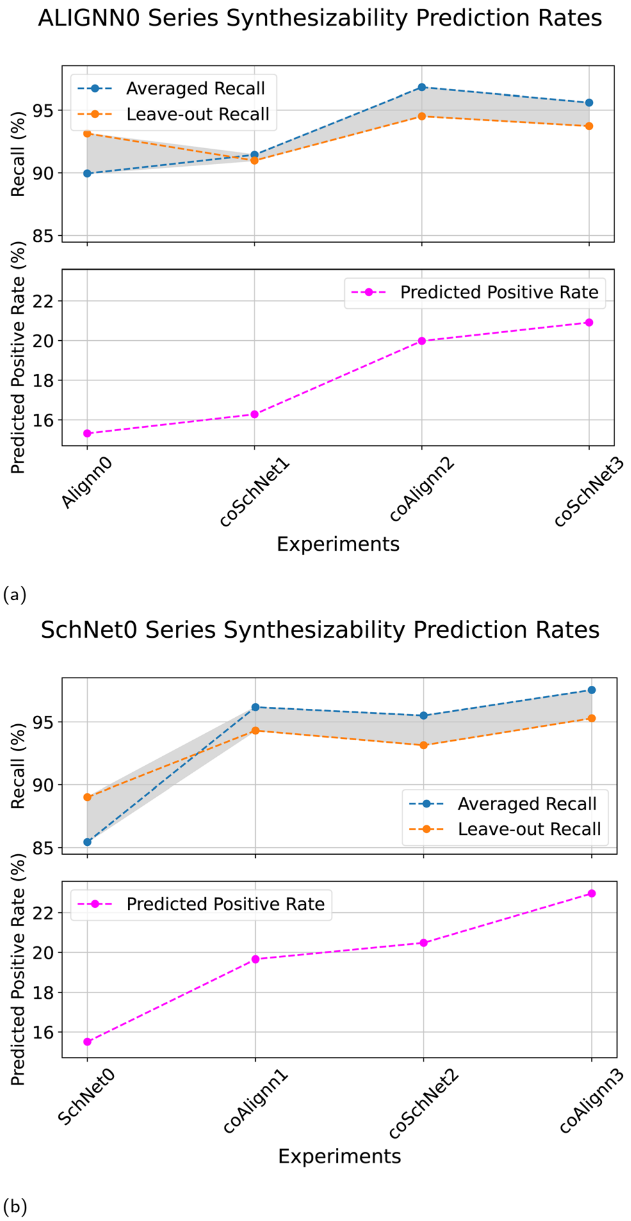


Fig. 2 Recall progression per iteration for both co-training series. The first series (a) starts by training a base PU learner with an ALIGNN classifier, whereas the second series (b) begins with a SchNet classifier.

recall range to justify it. This means that the model is more likely to classify a theoretical crystal as synthesizable, without improving its understanding of synthesizability. This is akin to over-fitting, when additional learning steps do not yield better validation results. These factors indicate that iteration ‘2’ is optimal. Consequently, we omit the third iteration and use the results from iteration ‘2’ as the source for synthesizability labels.

The synthesizability scores provided for the unlabeled data is the actual goal of this PU learning task. The distribution of these scores, alongside a large recall rate, provide a sense of the performance quality. A model that marks almost all crystals as synthesizable would have a high recall but could not distinguish the two classes from each other. Fig. 3 and 4 show the distributions of synthesizability scores at iteration ‘0’ and iteration ‘2’ of co-training, respectively. Despite high recall values, the PU learners mark only about 20% of the unlabeled data as synthesizable. The synthesizability scores for the intermediate iterations can be found in the ESI.†

In the final step of co-training, the scores from iteration ‘2’ are averaged and final labels are predicted *via* a cutoff of 0.5. This yields the final labels to for training the synthesizability predictor. The recall range is now [95–97]% and 21% of the unlabeled data are predicted to be synthesizable, see Fig. 5. Of course, all experimental data, including the ∼3% that were misclassified as unsynthesizable, are labeled as positive for training the synthesizability predictor.

Next, we examine the synthesizability score for the unlabeled data and its relationship to stability. Crystals with energy more than 1 eV above the convex hull are considered highly unstable and unlikely to be synthesizable. As shown in Fig. 6, the majority of our dataset consists of stable materials, indicating that our synthesizability predictions largely exclude unstable data. Furthermore, Fig. 6 reveals that unstable crystals are 2.5 times less likely to be classified as synthesizable than as non-synthesizable. However, among all crystals with energy less than 1 eV above the hull, only about 21% are classified as synthesizable. Additionally, we observe a sharp decline in energy above the hull when the synthesizability score increases slightly from zero (contour line). Conversely, materials that are confidently predicted to be synthesizable exhibit an increase in energy above the hull.

While one might expect stability to correlate directly with higher synthesizability scores, this trend is not strongly demonstrated here, likely due to the limited number of unstable crystals in our dataset. Although stability plays a significant role in synthesizability, it is not expected to be the sole determining factor. It is also important to acknowledge the inherent limitations of DFT, such as finite temperature effects

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

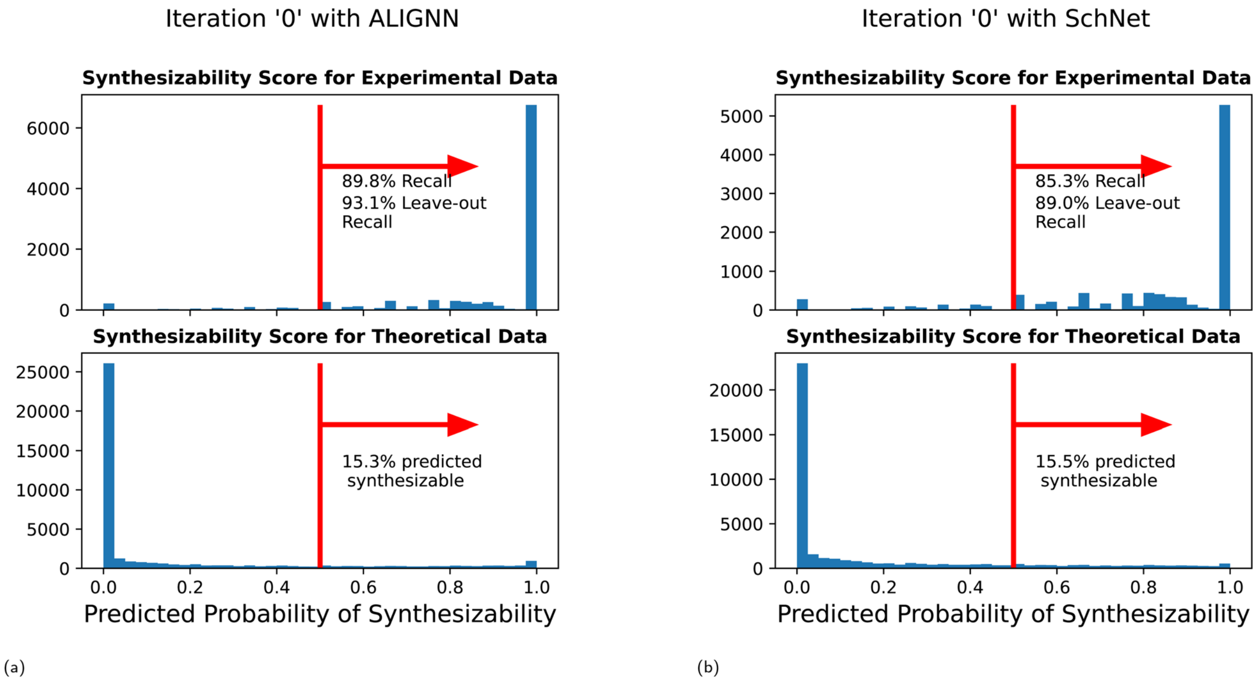


Fig. 3 Synthesizability score distribution for Iteration '0' for the (a) ALIGNN0 series and (b) SchNet0 series. The top half displays the experimental data. Ideally, a single peak at 1 would be observed, as these materials have been synthesized. However, a few intermediate scores (and some below 0.5) are present, which indicates that the model has not yet fully learned the target. The bottom half shows that most data points are classified as 0. However, a meaningful portion of the data (15% above the 0.5 cutoff) falls in the middle. Later iterations show a clear peak at 1.

and precision constraints, which may influence these observations.

In Fig. 7, we compare the energy above hull and formation energy for data with positive and negative labels. The left column presents the experimental data, while the right column corresponds to the actual task of distinguishing positive and negative classes within the unlabeled data. As expected, we observe a clustering of positive data around lower values of

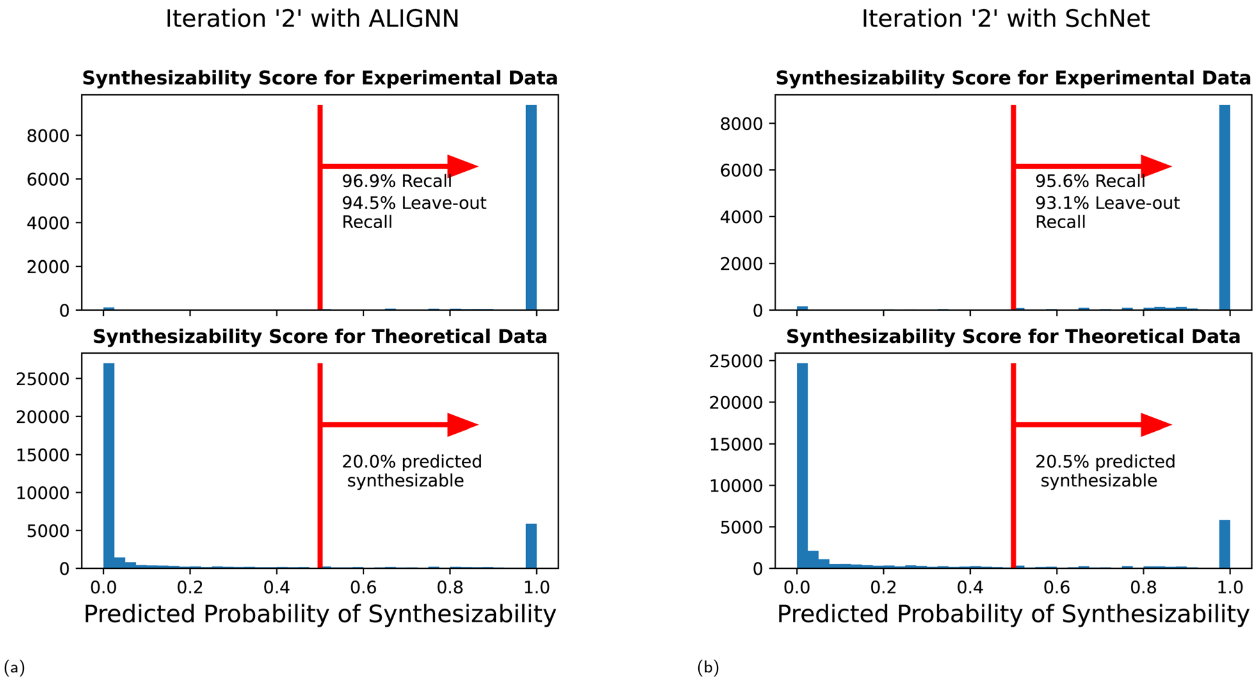


Fig. 4 Synthesizability score distribution for Iteration '2' for the (a) ALIGNN0 series and (b) SchNet0 series. We observe a bimodal distribution emerging in the unlabeled data, while the experimental data maintains a single peak. The bimodality results from the co-training mechanism, where positively predicted materials are integrated into the positive set of the training data.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

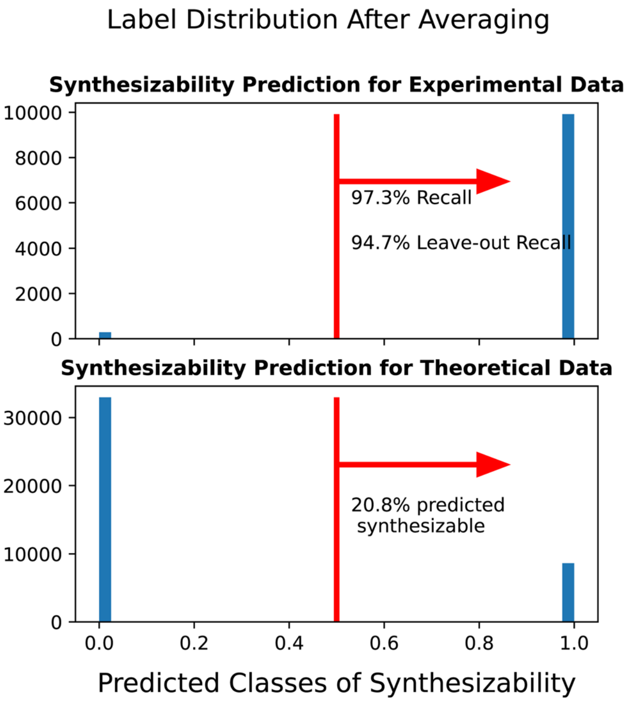


Fig. 5 Label distribution after averaging scores. The averaged scores were converted to labels {0, 1} using an unbiased threshold of 0.5. A label of 1 is assigned to materials that are predicted to be synthesizable. The plot at the top shows that most experimental data has been correctly classified as synthesizable.

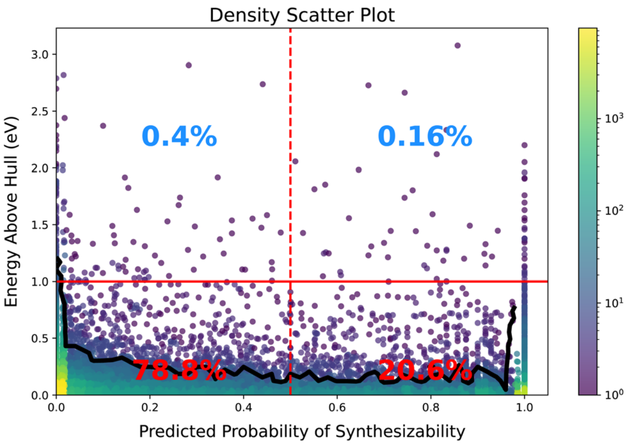


Fig. 6 Density scatter plot of energy above hull *vs.* the synthesizability score for the unlabeled data. The logarithmic color bar on the side indicates density map. The black line shows a single contour line.

energy above hull, without any distinct density peaks in formation energy. This aligns with our expectations, as stability (and, by extension, synthesizability) is influenced more by relative energy states than by absolute energy values.

2.3 Ground truth evaluation

2.3.1 Test-sets construction. Construction of a suitable test-set is required to report any type error or merit measure in machine learning; more-so when predicting out-of-distribution data is concerned.[25] In their paper establishing the bagging PU learning method,[31] Mordelet and Vert use a different test-set at each run. After all runs have been executed, the average label predicted for each datum as part of the test-set is taken as the probability of it belonging to the positive class. A threshold is then applied to this probability to determine the label of the datum, and thus, the error criteria. This does not lead to 'data leakage' as no model is tested on the data it has been trained on. This dynamic test-set also lends itself well to co-training, as it does not take away valuable data permanently from the growing train-sets of further iterations.

On the other hand, using a leave-out test-set in the common practice in materials informatics. At the very least, a leave-out test-set would provide a more comparable evaluation with similar works in this topic.

Ultimately, the goal of a test error is to approximate the expected test error. By using both test sets, we will have two values for recall. That means more information about the model's performance. We chose a leave-out test set with 5% of the positive data for all the runs. For the dynamic test set, 10% of the positive data is chosen at each run of the PU learning.

2.3.2 Ground truth in PU learning. Recall is the typical measure for evaluating PU learning tasks. Due to the unlabeled data, this is not the most reliable measure. Recall only tells us how much of the known positive samples were classified correctly. The assumption is that the positive data are sampled from an unknown distribution. Hence, the recall based on the labeled data should approximate a recall based on all the positive data. Yet, it would be beneficial to have some evidence, even if qualitative, that the recall solely based on the labeled data in fact approximates a recall based on all the data, the ground truth recall. To that end, we construct a new PU learning task with the goal of predicting classes of stability. As mentioned earlier, stability and synthesizability are related properties. If the recall values for labeled stability classes closely approximate the ground truth recall for all of the data, this suggests a similar behavior in the recall values for synthesizability.

We use the same dataset as before, now including the outliers of experimental structures with high energy above hull that were previously excluded. This adjustment retains data for learning the higher energy structure and provides a better benchmark for comparison with previous works in synthesizability that used the outliers. These data points were classified into positive (stable) and negative (unstable) classes based on a cutoff in energy above the convex hull; for details please see ESI.† The key difference is that, unlike a real PU learning task, all the positive and negative labels are available for evaluation post training. A random subset of the positive class, with the same number of data points as the original experimental class, kept their positive label. We then hid the label of the remaining data to manufacture a PU learning scenario. The models were trained on the stability PU data using the same code as the synthesizability task. Having access to all the labels, we could estimate the ground truth recall value and compare it with the recall values produces by the two test sets.

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

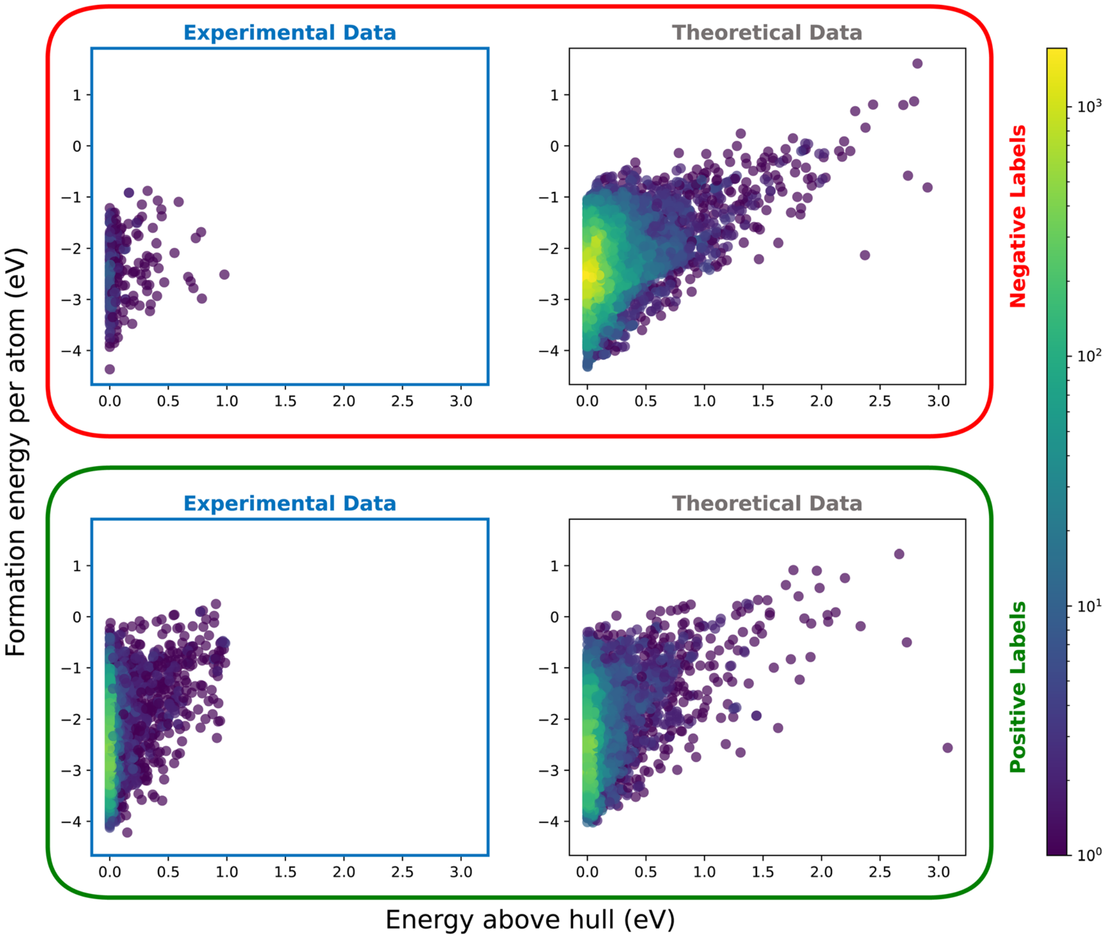
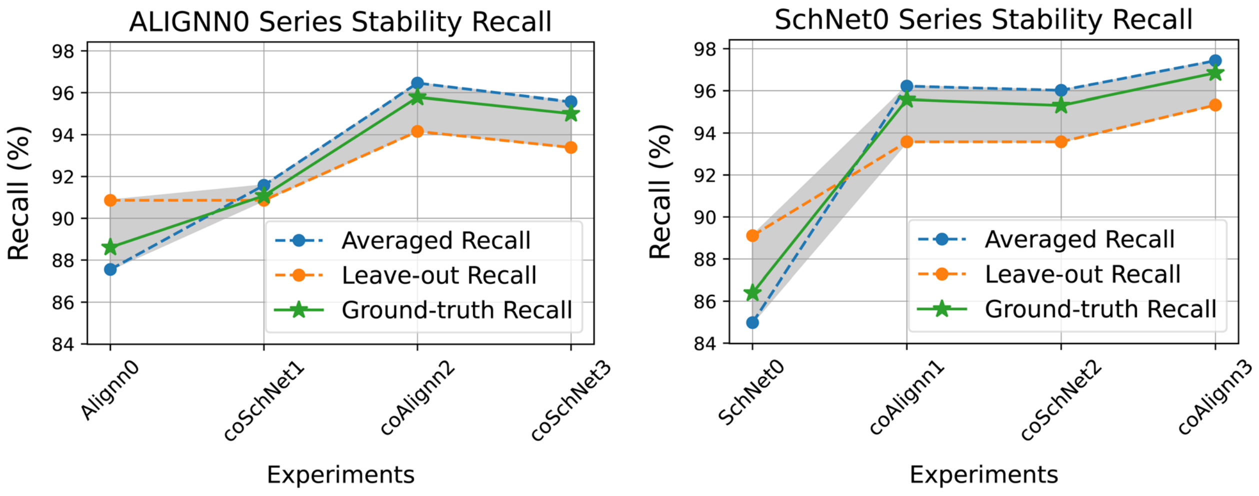


Fig. 7 Distribution of formation energy *versus* energy above hull for experimental data (left) and theoretical data (right), separated by predicted labels. The figure illustrates the expected clustering of positively labeled data around lower values of energy above hull while no distinct density peaks are observed with respect to formation energy.

As shown in Fig. 8, the recall values produced by both test-sets closely approximate the ground truth recall, confirming the reliability of using recall for evaluating the model's performance. In both co-training series, the leave-out recall value starts more optimistic than the ground truth, especially when high-energy experimental outliers are included in the PU



(a) (b)

Fig. 8 Ground truth recall progression per iteration for stability classes for the first (a) and second (b) series.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

learning. This optimistic recall was the reported recall value in the previous PU learning studies predicting synthesizability.[14,23] From iteration '1' of co-training the order flips and the dynamic test set becomes too optimistic. While there is no guarantee the ground-truth will always be found in the range between the two values, Fig. 8 illustrates why using both test-sets is worthwhile rather than just keeping one.

2.4 Predicting synthesizability

The final step in this work involves training a synthesizability predictor using reliable labels generated through co-training. Since materials databases apply different criteria for data inclusion, they exhibit varied data distributions. Consequently, achieving optimal performance on a specific test set is insufficient. It is crucial to avoid bias toward the Materials Project data distribution, which was used to train the model. To mitigate this, we applied regularization techniques during training to prevent overfitting. This was not particularly important in the PU runs, where classifier instability enhances the bagging process by introducing variability.[31] In the final step, however, the model needs to generalize well to data distributions not seen during training, while still maintaining good performance on the test-set.

We selected SchNet as our classifier and achieved good results, though other classifiers like ALIGNN can also be trained using the same labels. Detailed training parameters are available in the METHODS section. The pretrained model is accessible in our repository (https://github.com/BAMeScience/SynCoTrainMP).

The trained model reached 90.5% accuracy on a test set comprising 5180 data points. To further evaluate the model's performance, we analyzed the synthesizability predictions for three additional datasets, focusing exclusively on oxides. These datasets originate from other sources than our training data and consequently exhibit different biases.[35] First, we examined theoretical oxides from the Open Quantum Materials Database (OQMD),[37] downloaded *via* the Jarvis Python package,[38] after filtering out any crystals already present in Materials Project's experimental data, leaving 23 056 theoretical oxides. Second, we analyzed 14 095 oxide crystals from the WBM dataset,[39] which were generated using random sampling of elements in Materials Project structures, with chemical similarity safeguards based on ICSD data.[34] We used the relaxed version of this dataset. Finally, we predicted the synthesizability of 6156 vanadium oxide crystals generated by iMatGen.[11] Fig. 9 compares the synthesizability scores of these datasets with the theoretical portion of the test set. All these crystal structures and their predicted synthesizability scores are available to download in our GitHub repository.

Over half of the theoretical test-set data shows a synthesizability score close to zero, as expected, since previously synthesized crystals have been excluded by Materials Project. In contrast, the OQMD data shows roughly twice the proportion of synthesizable crystals, which may result from differing inclusion criteria between Materials Project and OQMD. We still observe a peak near a score of 1, possibly indicating synthesized crystals not listed in Materials Project. The iMatGen data show the lowest synthesizability, with multiple peaks at low scores, reflecting the artificial nature of these generated structures, which are often less realistic. The WBM dataset scores higher on average, without significant peaks. Despite being artificially generated, the WBM data employed mechanisms like chemical similarity to avoid unstable crystals. As a result, we observe more novel crystals with ambiguous synthesizability predictions, with scores around 0.5, and no clear peaks close to 0 or 1.

2.5 Discussion

In modern material discovery, the first challenge is the abundance of choices. The number of materials that may exist is astronomical[40] and high-throughput methods cannot screen

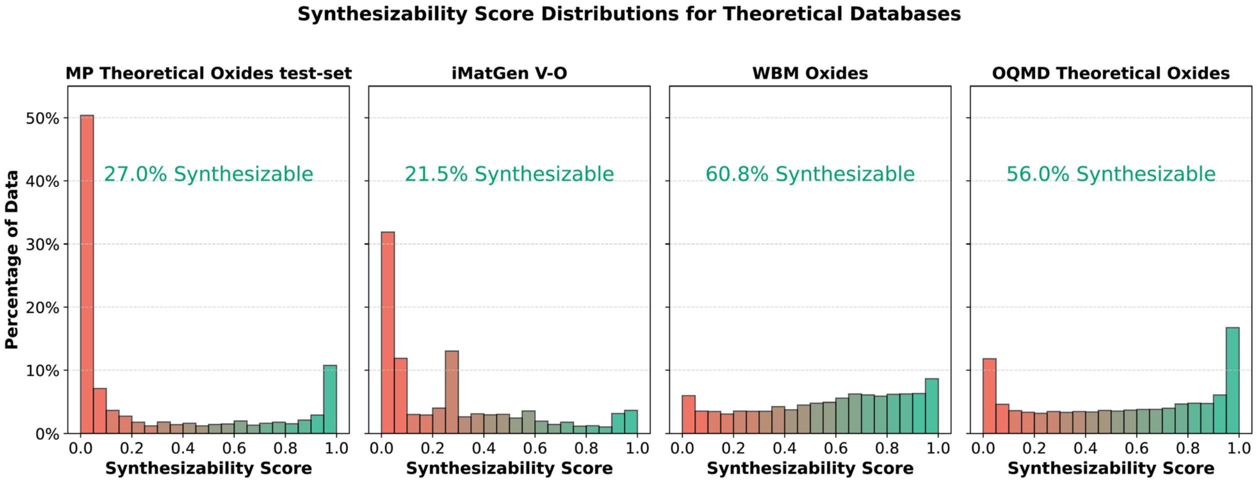


Fig. 9 Synthesizability probability distributions across theoretical databases, expressed as a percentage of each dataset. The first distribution represents the theoretical portion of the test set, selected from Materials Project data. The second distribution corresponds to vanadium oxide crystals generated by the iMatGen generative model. The third distribution shows data from the WBM dataset, followed by the fourth, which represents oxides from the OQMD database that are absent in Materials Project experimental data.

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

them all. The nebulous nature of the synthesizability question makes the development of a flawless model challenging. The goal, however, is not perfection. Based on the findings of this and other related studies, the majority of the unlabeled data is determined to be unsynthesizable. Energy calculations, while important, are not a good proxy for synthesizability. Filtering out even half of the unsynthesizable data through synthesizability prediction could save a significant amount of resources on simulations and synthesis attempts. We imagine that our tool can be used in the initial stages of materials discovery, to filter out the structures which are not likely to result in real materials.

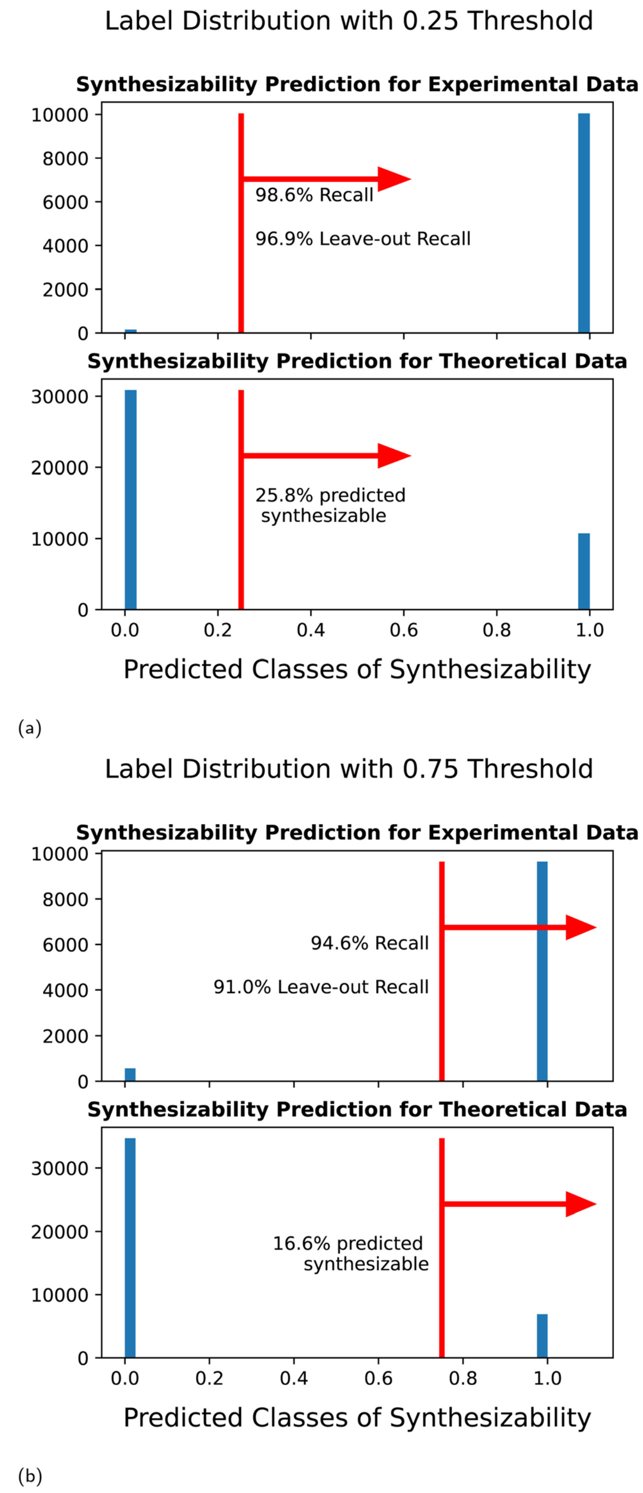


Fig. 10 Label distribution based on 0.25 and 0.75 classification thresholds at the end of co-training for the first (a) and second (b) series.

The decision thresholds of 0.5 and 0.75 were used as unbiased values for classification and class expansion. However, these thresholds are ultimately arbitrary and can be adjusted based on the specific goals and applications. In a more exploratory study, a looser threshold could be utilized to avoid overlooking potentially interesting novel structures. Conversely, a project operating with a tighter budget could employ a stricter threshold to save on resources. Label distributions based on a threshold of 0.25 and 0.75 are illustrated as examples in Fig. 10. When compared with the unbiased threshold of 0.5 shown in Fig. 5, a cutoff threshold of 0.25 is more lenient in classifying crystals as synthesizable. However, it only identifies 26% of the theoretical oxides in Materials Project as synthesizable, recognizing two thirds of the data as unsynthesizable. Conversely, a threshold of 0.75 results in a more stringent classification, with only 17% of the theoretical oxides meeting this threshold. And yet, these oxides are more likely to be synthesizable compared to those that did not meet the cut.

In this work we combined two different learners based on strong classifiers to reach a more reliable recall. New models for predicting materials properties are developed rapidly and materials data is growing. Combining different tools, instead creating one from scratch, is an untapped potential to learn more from the data already available in the materials space.

3 Methods

3.1 Co-training

The co-training algorithm used here was based on previous work in ref. 26 and 27. It is based on the idea that each data point can be described by distinct sets of descriptors, each of which are sufficient for learning the target. Consequently, two models learning from different views of the data can each gain knowledge that is inherently complementary. This is analogous to transfer learning; but the transfer happens between the knowledge gained from different views of the same data, rather than an auxiliary data source.

At each iteration, a base learner calculates a synthesizability score between 0 and 1 for both the unlabeled and experimental test data. To expand the positive class, unlabeled data points confidently classified as positive by the PU learner are selected. Here, we use a threshold of 0.75, rather than 0.5, to determine which unlabeled data points are added in the original positive class. After iteration ‘2’, the scores from both training series are averaged. The 0.5 cutoff determines the final label.

The base learner was changed from the original naïve Bayes classifier to a base PU learners equipped with convolutional neural networks. The different views of the data were achieved through the difference between the data encoding in the

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

classifiers, *i.e.*, ALIGNN and SchNet. Two parallel co-training series with altering classifiers were carried out accordingly.

3.2 PU learning

The algorithm of PU learning was established by Mordelet and Vert.[31] This method treats the unlabeled data as negative data, contaminated with positive data. PU learning performs best when this contamination is low.

In this work, two base PU learners were made through using two classifiers. In both cases, a complete bagging of PU learning took 60 runs. Note that the separate runs of PU learning are not referred to as iterations as each run is independent of the rest. This is not the case in co-training, where each iteration depends on the results produced by the previous iteration.

The training data at each PU learning run has a 1 : 1 ratio of positive and negative labels. The size of the training set increases after each co-training iteration, due to the expansion of the positive class. Each run of PU learning predicts a label, 0 or 1, for the data points that did not take part in the training phase of that run. After the 60 runs, these predictions are averaged for each data to produce the synthesizability score. This score is also referred to as the predicted probability of synthesizability. The cutoff thresholds of 0.5 and 0.75 are used to predict the labels and expand the positive class, respectively.

3.3 Neural networks architecture

The ALIGNN model was used according to instructions provided in its repository. We used the version 2023.10.01 of ALIGNN.

The SchNetPack model was originally designed for regression. To accommodate classification, a sigmoid non-linearity and a cutoff function were added to the final layer. We used the version 1.0.0.dev0 of this model.

3.4 The synthesizability predictor

The data labeling process utilized the averaged synthesizability scores produced in iteration '2' of co-training. A cutoff of 0.75 was applied for assigning positive labels, similar to the class expansion strategy, reducing the likelihood of training on uncertain labels.

During initial tests, the predictor displayed a tendency to overestimate the positive class, likely due to overfitting to the data distribution in the Materials Project. To mitigate this, several regularization steps were introduced. First, noise was added to the labels by randomly selecting 5% of the positive class and flipping their labels from 1 to 0. An equal number of negative class labels were also flipped from 0 to 1. This small amount of label noise helps regularize the model, preventing classifier's overconfidence in any class distribution.

Data augmentation was then employed, following a previously published method that showed significant improvements in predicting material properties. This approach perturbs atomic positions using Gaussian noise to generate slightly altered versions of the original data, which are used alongside the unperturbed data for training. This augmentation doubles the size of the training set.

The SchNet model was used as the primary synthesizability predictor, with additional regularization techniques enhancing its generalizability. A weighted loss function was employed, with a ratio of 0.45 : 0.55 for the positive and negative classes, respectively. This adjustment subtly discouraged over-prediction of the positive class, while maintaining model sensitivity.

Finally, to implement regularization during training, dropout layers were added to the model, with 10% dropout at the embedding layer and 20% at each convolutional layer. To manage the learning rate, a 'Cosine Annealing with Warm Restarts' scheduler was used, allowing it to cycle through phases, helping the model escape local minima early in training while converging effectively later on. Early stopping was also implemented to prevent overtraining.

3.5 Datasets

The experimental and theoretical data for co-training was queried from the Materials Project API,[6] database version 2023.11.1.

Open Quantum Materials Database (OQMD)[37] served as an external dataset that was not used in model training. However, they were downloaded through the Jarvis Python package[38] on 2023.12.12, which provides easy access to this data.

The WBM dataset[39] was made available through the Matbench Discovery[41] project through figshare.[42]

Data availability

Software code and all results presented in this paper, including intermediate PU-learning steps are available at: **https://doi.org/10.5281/zenodo.14411489**. Our method is implemented in a public software repository: **https://github.com/BAMeScience/SynCoTrainMP**. The software corresponding to this publication is available at: **https://github.com/BAMeScience/SynCoTrainMP/tree/0.0.3**. The input data for our algorithm and the code for downloading and processing it is available in our software repository.

Conflicts of interest

There are no conflicts to declare.

References

1 J. J. De Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton and J.-C. Zhao, *npj Comput. Mater.*, 2019, **5**, 41.
2 A. White, *MRS Bull.*, 2012, **37**, 715–716.
3 J. R. Rodgers and D. Cebon, *MRS Bull.*, 2006, **31**, 975–980.
4 L. Pauling, *J. Am. Chem. Soc.*, 1929, **51**, 1010–1026.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 March 2025. Downloaded on 2/25/2026 2:46:01 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

5 E. R. Antoniuk, G. Cheon, G. Wang, D. Bernstein, W. Cai and E. J. Reed, *npj Comput. Mater.*, 2023, **9**, 155.

6 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

7 J. George, D. Waroquiers, D. Di Stefano, G. Petretto, G. Rignanese and G. Hautier, *Angew. Chem.*, 2020, **132**, 7639–7645.

8 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, *Sci. Adv.*, 2016, **2**, e1600225.

9 T. F. T. Cerqueira, S. Lin, M. Amsler, S. Goedecker, S. Botti and M. A. L. Marques, *Chem. Mater.*, 2015, **27**, 4562–4573.

10 A. K. Singh, J. H. Montoya, J. M. Gregoire and K. A. Persson, *Nat. Commun.*, 2019, **10**, 443.

11 J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, *Matter*, 2019, **1**, 1370–1384.

12 Y. Wu, P. Lazic, G. Hautier, K. Persson and G. Ceder, *Energy Environ. Sci.*, 2012, **6**, 157–168.

13 C. J. Bartel, *J. Mater. Sci.*, 2022, **57**, 10475–10498.

14 J. Jang, G. H. Gu, J. Noh, J. Kim and Y. Jung, *J. Am. Chem. Soc.*, 2020, **142**, 18836–18843.

15 A. Lee, S. Sarker, J. E. Saal, L. Ward, C. Borg, A. Mehta and C. Wolverton, *Commun. Mater.*, 2022, **3**, 73.

16 K. Li and W. Chen, *Mater. Today Energy*, 2021, **20**, 100638.

17 M. Miao, Y. Sun, E. Zurek and H. Lin, *Nat. Rev. Chem*, 2020, **4**, 508–527.

18 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

19 F. Legrain, J. Carrete, A. van Roekeghem, G. K. Madsen and N. Mingo, *J. Phys. Chem. B*, 2018, **122**, 625–632.

20 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, *npj Comput. Mater.*, 2019, **5**, 62.

21 A. Davariashtiyani, Z. Kadkhodaie and S. Kadkhodaei, *Commun. Mater.*, 2021, **2**, 115.

22 M. Aykol, V. I. Hegde, L. Hung, S. Suram, P. Herring, C. Wolverton and J. S. Hummelshøj, *Nat. Commun.*, 2019, **10**, 2018.

23 G. H. Gu, J. Jang, J. Noh, A. Walsh and Y. Jung, *npj Comput. Mater.*, 2022, **8**, 71.

24 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.

25 K. Li, B. DeCost, K. Choudhary, M. Greenwood and J. Hattrick-Simpers, *npj Comput. Mater.*, 2023, **9**, 1–9.

26 A. Blum and T. Mitchell, *Proceedings of the eleventh annual conference on Computational learning theory*, Madison Wisconsin USA, 1998, pp. 92–100.

27 F. Denis, A. Laurent, R. Gilleron and M. Tommasi, *Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data*, 2003, vol. 8.

28 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 185.

29 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.

30 K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K.-R. Müller, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, year.

31 F. Mordelet and J.-P. Vert, *Pattern Recognit. Lett.*, 2014, **37**, 201–209.

32 R. Savkina and L. Khomenkova, *Oxide-Based Materials and Structures: Fundamentals and Applications*, CRC Press, 2020.

33 D. Waroquiers, X. Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Göbel, S. Schenk, P. Degelmann, R. André, R. Glaum and G. Hautier, *Chem. Mater.*, 2017, **29**, 8346–8360.

34 M. Hellenbrandt, *Crystallogr. Rev.*, 2004, **10**, 17–22.

35 A. Davariashtiyani, B. Wang, S. Hajinazar, E. Zurek and S. Kadkhodaei, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 040501.

36 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

37 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, *npj Comput. Mater.*, 2015, **1**, 15010.

38 K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, *npj Comput. Mater.*, 2020, **6**, 173.

39 H.-C. Wang, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2021, **7**, 1–9.

40 D. Davies, K. Butler, A. Jackson, A. Morris, J. Frost, J. Skelton and A. Walsh, *Chem*, 2016, **1**, 617–627.

41 J. Riebesell, R. Goodall, P. Benner, Y. Chiang, B. Deng, A. Lee, A. Jain and K. Persson, Matbench Discovery: An evaluation framework for machine learning crystal stability prediction, 2023, available at, **https://janosh.github.io/matbench-discovery**.

42 Matbench Discovery v1.0.0, 2023, available at, **https://figshare.com/articles/dataset/Matbench_Discovery_v1_0_0/22715158/6**.

© 2025 The Author(s). Published by the Royal Society of Chemistry