

ARTICLE OPEN



Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions

Evan R. Antoniuk^{1,2}, Gowoon Cheon^{3,4}, George Wang^{5,6,7}, Daniel Bernstein⁸, William Cai⁸ and Evan J. Reed⁸

Reliably identifying synthesizable inorganic crystalline materials is an unsolved challenge required for realizing autonomous materials discovery. In this work, we develop a deep learning synthesizability model (*SynthNN*) that leverages the entire space of synthesized inorganic chemical compositions. By reformulating material discovery as a synthesizability classification task, *SynthNN* identifies synthesizable materials with 7× higher precision than with DFT-calculated formation energies. In a head-to-head material discovery comparison against 20 expert material scientists, *SynthNN* outperforms all experts, achieves 1.5× higher precision and completes the task five orders of magnitude faster than the best human expert. Remarkably, without any prior chemical knowledge, our experiments indicate that *SynthNN* learns the chemical principles of charge-balancing, chemical family relationships and ionicity, and utilizes these principles to generate synthesizability predictions. The development of *SynthNN* will allow for synthesizability constraints to be seamlessly integrated into computational material screening workflows to increase their reliability for identifying synthetically accessible materials.

npj Computational Materials (2023)9:155; <https://doi.org/10.1038/s41524-023-01114-4>

INTRODUCTION

Throughout the history of modern science, the discovery of novel materials with technologically desirable properties has resulted in rapid scientific innovation. The first step in the discovery of any new material is to identify a novel chemical composition that is synthesizable, which we here define to be a material that is synthetically accessible through current synthetic capabilities, but may or may not have been synthesized yet. Our ability to develop new materials and technologies is therefore dependent on our ability to efficiently search through the entirety of chemical space to identify synthesizable materials for further investigation.

For the purposes of this work, we refer to *synthesized* materials as the set of all materials that have had their synthesis details reported in the literature or are naturally occurring. Predicting synthesized materials is of little interest since this task can be trivially accomplished by searching through existing materials databases^{1,2}. Instead, this work explores the question of whether it is possible to develop a method for predicting the *synthesizability* of inorganic crystalline materials, regardless of whether or not that material has been synthesized yet.

Whereas organic molecules can often be synthesized through a sequence of well-established chemical reactions³, the targeted synthesis of crystalline inorganic materials is complicated by the lack of well-understood reaction mechanisms⁴. Instead, specific inorganic materials can be preferentially synthesized through selecting reactants that provide thermodynamic or kinetic stabilization of the product, choosing reaction pathways that minimize unwanted side-products, and/or encouraging selective nucleation of the target material^{4–6}. However, the decision to synthesize a material also depends on a wide range of non-physical considerations including the cost of the reactants, the availability of equipment required for the synthesis, and the human-perceived importance of the final product. As a result,

synthesizability cannot be predicted based on thermodynamical or kinetic constraints alone. The final decision to pursue the synthesis of a target inorganic material has traditionally been the responsibility of expert solid-state chemists who specialize in specific synthetic techniques or classes of materials. The careful consideration of these experts minimizes the chance of an unsuccessful synthetic effort, but does not allow for rapid exploration of inorganic material space.

Due to the lack of a generalizable synthesizability principle for inorganic materials, the enforcement of a charge-balancing criteria is a commonly employed proxy for synthesizability (Fig. 1a)^{7–9}. This computationally inexpensive approach filters out materials that do not have a net neutral ionic charge for any of the elements' common oxidation states. Despite the chemically motivated nature of this approach, we find that charge-balancing cannot accurately predict synthesizable inorganic materials. Among all inorganic materials that have already been synthesized, only 37% can be charge-balanced according to common oxidation states (Supplementary Table 3) Remarkably, even among all ionic binary cesium compounds which are typically considered to be governed by highly ionic bonds, only 23% of known compounds are charge balanced (Supplementary Table 5). The poor performance of this charge-balancing approach is likely due to the inflexibility of the charge neutrality constraint, which cannot account for the different bonding environments that are present among different classes of materials such as metallic alloys, covalent materials, or ionic solids.

A wide array of ab-initio and machine learning methods have been developed to aid in the discovery of synthesizable materials. One commonly employed approach utilizes density-functional theory (DFT) to calculate the formation energy of a material's crystal structure with respect to the most stable phase in the same chemical space as the material of interest². This approach assumes

¹Materials Science Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA. ²Department of Chemistry, Stanford University, Stanford, CA, USA. ³Department of Applied Physics, Stanford University, Stanford, CA, USA. ⁴Google Research, Mountain View, CA, USA. ⁵Department of Physics, Stanford University, Stanford, CA, USA. ⁶Department of Mathematics, Stanford University, Stanford, CA, USA. ⁷Department of Computer Science, Stanford University, Stanford, CA, USA. ⁸Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA. ✉email: antoniuk1@llnl.gov

that synthesizable materials will not have any thermodynamically stable decomposition products. However, due to a failure to account for kinetic stabilization, this approach has previously been demonstrated to be unable to distinguish synthesizable materials from those that have yet to be synthesized and only captures 50% of synthesized inorganic crystalline materials^{10–12}. A wide range of machine learning-based composition models for predicting thermodynamic stability have also been developed as a means for assessing composition synthesizability (Fig. 1a)^{13–17}. Formation energy calculations have also been combined with the experimental discovery timeline to generate a materials stability network that can produce synthesizability predictions. More recently, several machine-learning based methods have been developed to predict the synthesizability of a material from its crystal structure^{11,18,19}. However, these methods require the atomic structure as an input, which is typically not known for materials that have yet to be discovered. To enable predictions on materials for which the crystal structure is not known, several composition-based material representations have been developed, predominantly for the task of material property prediction^{13,16,17}. Composition-based representations enable material discovery across the entirety of chemical composition space, but are unable to differentiate between different crystal structures of the same chemical composition. Finally, numerous works have utilized data-mining to extract inorganic material synthesis recipes from the literature^{20,21}. These methods can prescribe a synthesis recipe for a hypothetical material, but do not allow for an assessment of the synthesizability of the hypothetical material^{22,23}.

In this work, we develop a deep-learning classification model (that we call *SynthNN*) to directly predict the synthesizability of inorganic chemical formulas without requiring any structural information. We accomplish this goal by training *SynthNN* on a database of chemical formulas consisting of previously synthesized crystalline inorganic materials that has been augmented with artificially generated unsynthesized materials. *SynthNN* offers numerous advantages over previous methods for identifying synthesizable materials. Whereas expert synthetic chemists typically specialize in a specific chemical domain of a few hundred materials, this approach generates predictions that are informed by the entire spectrum of previously synthesized materials. Additionally, since this method trains directly on the database of all synthesized materials (rather than employing proxy metrics such as thermodynamic stability or charge-balancing), this approach also eliminates questions of how well these metrics can describe synthesizability. Rather, *SynthNN* learns the optimal set of descriptors for predicting synthesizability directly from the database of all synthesized materials, allowing it to better capture the complex array of factors that influence synthesizability. Finally, this method is computationally efficient enough to enable screening through billions of candidate materials. Since *SynthNN* can be seamlessly integrated with Materials Screening or Inverse Design workflows (Fig. 1c, d), the development of *SynthNN* serves to greatly improve the success rate and reliability of computational material discovery efforts by ensuring that the candidate materials discovered through these efforts are synthetically accessible.

RESULTS

Model development

One of the main challenges in predicting the synthesizability of crystalline inorganic materials lies in the lack of a generalizable understanding of what factors contribute to synthesizability. Although charge-balancing and thermodynamic stability are likely to play a role in the likelihood that a material has been synthesized, these features alone are unlikely to serve as a complete set of descriptors predicting synthesizability. To account

Table 1. Model hyperparameters used in training *SynthNN*.

Hyperparameter Name	Range of sampled values
Number of hidden units in hidden layer #1:	[30,40,50,60,80]
Number of hidden units in hidden layer #2:	[30,40,50,60,80]
Adam optimizer learning rate:	$[2 \times 10^{-2}, 5 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-4}, 2 \times 10^{-4}]$
Number of initial supervised training steps	$[2 \times 10^4, 4 \times 10^4, 6 \times 10^4, 8 \times 10^4, 1 \times 10^5]$
Batch size	[512,1024]
Atomic Embedding Dimension (M)	[2,6,10,15,20,25,30,35]
Ratio of Synthesized:Unsynthesized Training Examples (N_{synth})	[5,10,15,20]

for this challenge, we adopt a framework called *atom2vec*, which represents each chemical formula by a learned atom embedding matrix that is optimized alongside all other parameters of the neural network (Fig. 1b, Methods)^{7,24}. In this manner, *atom2vec* learns an optimal representation of chemical formulas directly from the distribution of previously synthesized materials. The dimensionality of this representation is treated as a hyperparameter whose value is set prior to model training (see Methods, Table 1). Notably, this approach does not require any assumptions about what factors influence synthesizability or what metrics may be used as proxies for synthesizability, such as charge balancing. The chemistry of synthesizability is entirely learned from the data of all experimentally realized materials.

The synthesizable inorganic materials that *SynthNN* is trained on are extracted from the Inorganic Crystal Structure Database (ICSD)²⁵. This database represents a nearly complete history of all crystalline inorganic materials that have been reported to be synthesized in the scientific literature and have been structurally characterized (see Methods). On the other hand, unsuccessful syntheses are not typically reported in the scientific literature. We treat this lack of recorded data on unsynthesizable materials by creating a Synthesizability Dataset that is augmented with artificially-generated unsynthesized materials. It is important to note that some of these artificially-generated materials could be synthesizable, but are absent from the ICSD database or have yet to be synthesized. Definitely labeling a material as unsynthesizable is potentially problematic since the ongoing development of synthetic methodologies may enable the synthesis of previously unsynthesizable materials. To account for this incomplete labeling of the artificially generated examples, we develop a semi-supervised learning approach that treats unsynthesized materials as unlabeled data and probabilistically reweights these materials according to the likelihood that they may be synthesizable (see Methods)²⁶. The ratio of artificially generated formulas to synthesized formulas used in training is a model hyperparameter that we refer to as N_{synth} (see Supplementary Note 1).

SynthNN therefore fits into a broader category of positive-unlabeled (PU) learning algorithms. Recently, PU learning approaches have been adopted in materials science to handle the large amount of unlabeled material data that exists because of the tiny fraction of chemical space that has been experimentally explored. A transductive bagging support vector machine approach has been previously used for predicting the synthesizability of crystals and the discovery of synthesizable MXenes^{11,27}. However, the PU learning approach in the present work most closely resembles the approach of Cheon et al., whereby unlabeled examples are class-weighted according to their likelihood of synthesizability²⁶.

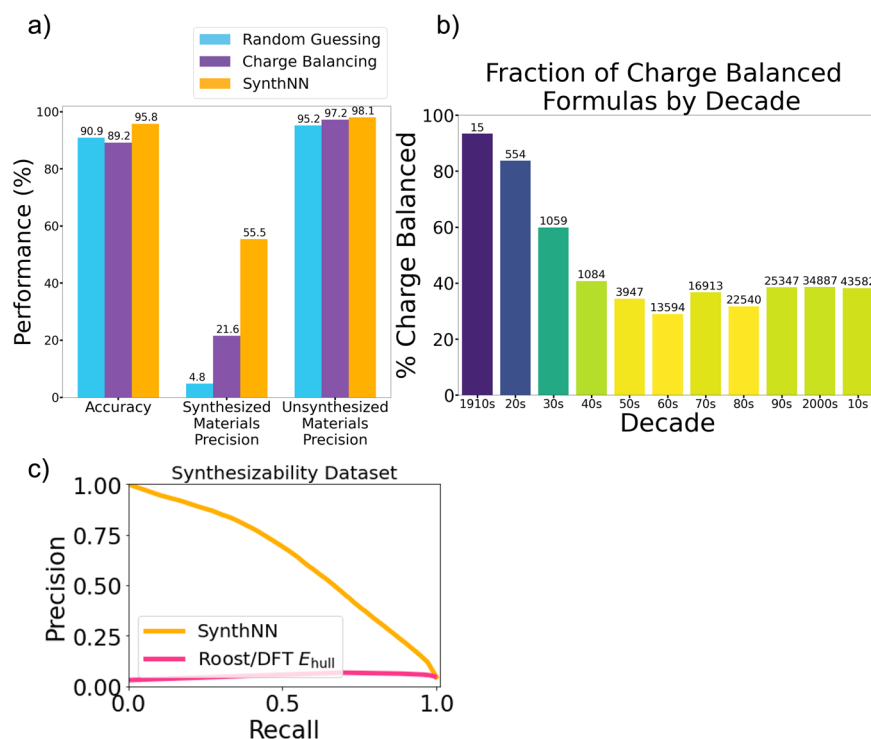


Fig. 2 Benchmarking *SynthNN* against other computational methods. Additional performance metric comparisons between *SynthNN*, charge-balancing and Roost are provided in Supplementary Table 4. **a** Performance of the *SynthNN* model on a test set with a 20:1 ratio of unsynthesized:synthesized materials ($N_{\text{synth}} = 20$), containing 2410 synthesized materials and 48,199 unsynthesized materials. We benchmark the performance of this model against a random guessing baseline and a charge balancing baseline (predicting a material to be synthesizable if it is charge balanced). The performance shown for the *SynthNN* model and charge-balancing is evaluated only on the test set. The random guessing baseline is taken to be the expected performance if randomly predicting 1/21 of all materials to be synthesizable in the full synthesizability dataset with $N_{\text{synth}} = 20$. **b** Fraction of all unique binary, ternary, and quaternary ICSD compounds that are charge-balanced, plotted for the decade that the materials were first synthesized. The number above the bar indicates the total number of materials that are listed in the ICSD as having been synthesized in that decade. Duplicate formulas are removed when calculating the fraction of materials and the listed total number of materials. Charge-balancing is determined according to the oxidation states listed in Supplementary Table 3. **c** Precision-recall curve comparison between *SynthNN* and Roost¹⁷ for predicting the synthesizability of materials in the entire Synthesizability Dataset with $N_{\text{synth}} = 20$.

Benchmarking against computational methods

The performance of *SynthNN* is shown in Fig. 2a alongside random guessing and charge-balancing baselines. After model training, we calculate standard performance metrics by treating synthesized materials and artificially generated unsynthesized materials as positive and negative examples, respectively. This choice results in the positive class precision shown in Fig. 1a to be lower than the true model precision since synthesizable, but unsynthesized materials will be incorrectly treated as false positive predictions (see Methods). The random guessing baseline corresponds to the expected performance if one were to make random predictions weighted by the class imbalance, whereas the charge balancing approach simply predicts a material to be synthesizable only if it is charge balanced according to the common oxidation states listed in Supplementary Table 3. Although the performance metrics in Fig. 2 are intended to provide an intuitive comparison of model performance, PU learning algorithms are most commonly evaluated based on the F1-score (shown in Supplementary Table 4)²⁸.

Separating the model performance into class-specific precision provides interesting insights into each of these approaches for predicting synthesizability. Both *SynthNN* and the charge-balancing model perform similarly well at detecting the artificially-generated unsynthesized materials in our dataset. The ability of the charge-balancing model to predict these unsynthesized materials with a high precision is likely a direct consequence of our choice to use randomly generated atomic coefficients to

generate the unsynthesized formulas. Generating chemical formulas with random coefficients is unlikely to yield a formula with a net neutral oxidation state, which allows the charge-balancing method to consistently identify these unsynthesized formulas. However, we observe considerable differences for the case of synthesizable formulas. We find that our *SynthNN* model is able to detect synthesized materials with a precision that is 2.6x higher than charge-balancing and 12x better than random guessing (Fig. 2a). It is interesting to note that the predictive accuracy of charge-balancing has significantly decreased over time (Fig. 2b). This likely reflects that our ability to synthesize complex stoichiometries has vastly outpaced the predictive ability of a simple charge-balancing approach, which further highlights the need for a more complex synthesizability predictor that accounts for a broader range of synthesizability considerations.

Predicting the thermodynamic stability of a material with DFT has been widely adopted as a standard approach for computationally discovering synthesizable materials^{14,16,17,29}. As a means of further benchmarking *SynthNN*, we compare the synthesizability predictions of *SynthNN* to approaches based on DFT calculated formation energies (Fig. 2c). In a real-world material discovery problem, the general task is to search across broad regions of chemical space to accurately identify synthesizable materials. An ideal direct comparison of material discovery based on *SynthNN* and DFT would therefore involve identifying the synthesized materials from a random sample of chemical composition space. However, DFT requires a material's crystal

structure as input. Although the crystal structure of a given composition can be computationally predicted through methods such as ab-initio random structure searching (AIRSS), it is computationally infeasible to predict the crystal structure for even hundreds of unsynthesized chemical compositions^{30–32}. Existing materials databases are also not a suitable testbed for such a material discovery task since they contain an unrealistically high fraction of synthesized materials and do not uniformly sample chemical composition space. The high proportion of synthesized materials in these materials databases are expected to greatly underestimate the false positive rate compared to a realistic material discovery setting.

We account for these challenges by instead benchmarking *SynthNN* against a composition-based machine learning model, Roost¹⁷, to act as a surrogate model for DFT calculations of energy above the convex hull. Notably, Roost was recently shown to outperform previous machine learned models for predicting formation energies and achieved an accuracy that approaches the DFT error, relative to experiment²⁹. We first train Roost on DFT-calculated energy above the convex hull values from the Materials Project database, where the value is taken from the lowest energy polymorph of each composition². Notably, 37,538 out of 79,533 of the compositions in this dataset of energy above the convex hull values are present in the ICSD, and thus overlap with the synthesized examples in our Synthesizability dataset. This re-trained Roost model achieves a mean absolute error of 0.063 eV across the entire dataset, which is nearly identical to 0.06 eV mean absolute error of a previously reported Roost model trained on the Materials Project²⁹. Following training, we use Roost to predict the energy above the convex hull values for all entries in the Synthesizability Dataset. A material is predicted to be synthesizable by Roost if the predicted energy above the hull is below a cutoff value, $E_{\text{hull,cutoff}}$, where $E_{\text{hull,cutoff}}$ is evaluated at {0eV, 0.05eV, 0.1eV, 0.2eV, 0.3eV, ..., 1eV}. We construct a precision-recall curve for Roost by evaluating its precision and recall across these various $E_{\text{hull,cutoff}}$ values (Fig. 2c).

Roost achieves a maximum F1-score of 0.12 when using $E_{\text{hull,cutoff}} = 0.05\text{eV}$. At this $E_{\text{hull,cutoff}}$, Roost achieves a recall of 69% and a precision of 6.8%. At the same recall of 69%, *SynthNN* achieves a precision of 46.6%, nearly 7× higher than that of Roost. Although both *SynthNN* and Roost are composition-based ML models, the main difference between these approaches is that Roost is trained on DFT-calculated energy above the convex hull values, whereas *SynthNN* is trained as a synthesizability classifier. Based on the 7× higher precision achieved by *SynthNN*, we conclude that reformulating the material discovery problem as a synthesizability classification task, rather than an energy above the convex hull regression task, can be a powerful strategy towards improving the accuracy for discovering novel materials. The performance of Roost could likely be improved by training on a hypothetical dataset of DFT calculated formation energies with greater chemical diversity and a larger fraction of unstable structures than is available in the Materials Project. However, creating such a dataset is hindered by the computationally expensive challenge of finding the most stable crystal structure for any given chemical formula. On the other hand, simply training *SynthNN* on the growing list of synthesized chemical compositions allows for greater chemical diversity to be incorporated into the training data without incurring additional computational expense.

Finally, it is interesting to note that increasing the number of artificially generated unsynthesized materials in the training set of *SynthNN* significantly improves model performance (Supplementary Tables 1–2). Generating these unsynthesized materials works analogously to data generation techniques that have been employed for the task of image classification^{33–35}. For the task of classifying images, generating new artificial images to increase the amount of training data has been shown to be effective in improving image recognition accuracy³⁴. In an analogous fashion,

including artificially generated unsynthesized materials in training may help to better inform our model about how synthesizability is impacted by the atom types, atomic coefficients, and the number of different atom types in the chemical formula. Including artificially generated formulas also exposes our model to a much broader chemical space beyond what is represented if we only trained our model on synthesized materials. The results in Supplementary Tables 1–2 suggest that this data augmentation strategy improves *SynthNN*'s precision at identifying promising regions of chemical space for future material discovery efforts.

Benchmarking against human experts

Although the model performance shown in Fig. 2 establishes *SynthNN* to outperform prior computational methods for predicting synthesizability, realizing an acceleration in the rate of material discovery is dependent on the ability of *SynthNN* to outperform human experts. Since pursuing material synthesis requires considerable resources, we anticipate that *SynthNN* will only see adoption if it can offer higher precision synthesizability predictions than human expert chemists. To compare the performance of *SynthNN* against expert chemists, we create a Synthesizability Quiz of 100 formulas by randomly sampling 91 unsynthesized formulas and 9 synthesized formulas from the Synthesizability Dataset (Supplementary Fig. 7). The high proportion of unsynthesized formulas was chosen to simulate a realistic materials discovery setting where synthesizable materials are expected to be rare (Supplementary Note 3). A total of 20 human participants were instructed to select the 9 materials that they think have been synthesized, whereas *SynthNN*'s predictions were taken to be the 9 materials that were predicted to be most likely to be synthesizable. Comparing the model performance of *SynthNN* to human experts in this manner provides quantification for the magnitude of improvement that artificial intelligence can provide for discovering new materials, while also providing a more relatable understanding of how well the *SynthNN* model performs.

Remarkably, *SynthNN* correctly predicts 6/9 synthesizable materials, whereas the best human expert only recovered 4/9 synthesizable materials. Although the Synthesizability Quiz contains only 100 examples, we note that *SynthNN*'s precision on this quiz (0.667) is comparable to the precision achieved on a validation set with the same proportion of synthesizable materials (precision of 0.665 for a recall of 0.600, shown in Supplementary Fig. 4). Furthermore, the average number of correct guesses achieved by *SynthNN* across 267 unique quizzes drawn without replacement from the test set was 5.84, which is significantly higher than the best human performance of 4.00. As has been observed in a wide variety of contexts^{36–39}, it is interesting to note that the aggregate response of all human experts yields predictions that are considerably more accurate than the average individual human response. Since chemists tend to specialize in specific domains of chemistry, this finding may highlight the significant improvement in the ability to predict synthesizability when the diverse domain expertise from multiple experts are independently aggregated.

To further probe the performance of *SynthNN* and the human experts, we explore how their predictions are affected by the complexity of the chemical formula (number of unique atoms, Fig. 3b) and the chemical family that the formula belongs to (Fig. 3c). Interestingly, we find that the performance of human experts significantly decreases for chemical formulas that contain d-block and f-block elements, whereas the performance of *SynthNN* is relatively constant across all chemical families (Fig. 3c). Chemical formulas from the f-block of the periodic table are commonly avoided due to their scarcity, radioactivity, and/or high cost^{40,41}. We speculate that this aversion to utilizing d- and f-block materials decreases the familiarity of human experts with these elements, which results in a decreased ability of human experts to identify

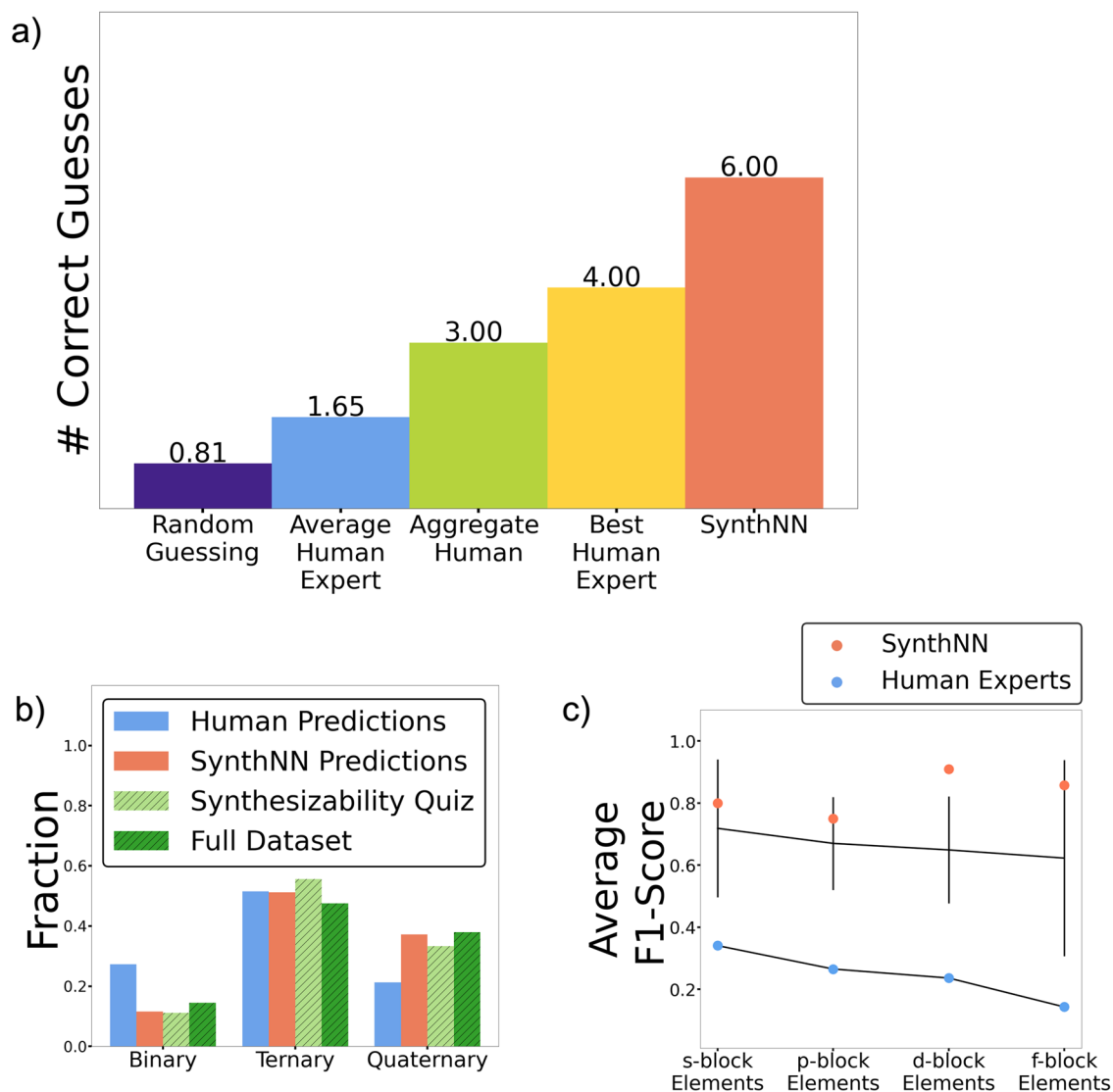


Fig. 3 Benchmarking *SynthNN* against human experts. **a** Number of the 9 synthesizable formulas in the Synthesizability Quiz that are correctly identified by a random guessing baseline (on average), the average human expert, the aggregate human response, the best performing human expert, and *SynthNN*. **b** Fraction of the human expert predictions on the Synthesizability Quiz that consist of binary, ternary, and quaternary compounds (blue). Fraction of *SynthNN* predictions on the Synthesizability Dataset test set that consist of binary, ternary, and quaternary compounds (orange). The true fraction of each compound type among the synthesized materials in the Synthesizability Quiz and the Synthesizability Dataset test set are shown for comparison. **c** Average F1-score of *SynthNN* and human experts on the Synthesizability Quiz, decomposed by periodic table blocks. A chemical formula is included in a block if any of its constituent elements belong to that block of the periodic table. The error bars of the *SynthNN* line provide the mean and standard deviation of the F1-score achieved by *SynthNN* across all quizzes (with 9 synthesized formulas and 91 unsynthesized formulas) generated from sampling without replacement from the test set of the Synthesizability Dataset.

synthesizable chemical formulas that contain them. It is also interesting to note that human experts vastly overestimate the likelihood that binary materials are synthesizable. Although binary materials only comprise 15% of all synthesized inorganic materials, binary materials accounted for 31% of the materials predicted by the human experts to be synthesizable. In comparison, since *SynthNN* has been trained on the distribution of all previously synthesized materials, it is able to generate synthesizability predictions that are well calibrated to match the distribution of formulas seen in known, synthesized materials (Fig. 3b). Finally, we note that the self-reported completion time for the Synthesizability Quiz was on the order of 30 min for the human experts, whereas *SynthNN* generates its predictions in a few milliseconds—corresponding to a 10^5 acceleration in prediction rate.

Predicting future materials

One of the most important and challenging tasks for any machine learning model is to make accurate predictions on examples that are drawn from a different distribution than the training set. This is particularly important for predicting synthesizability since new, innovative materials are expected to follow a significantly different distribution than the materials that have been previously synthesized. As a notable example, NaCl_3 has recently been shown to be synthesizable, in opposition to what would be predicted through charge-balancing, a DFT-calculated phase diagram, or traditional chemical intuition⁴². With the continued development of synthetic technologies, we expect that future material discovery efforts will increasingly move towards materials that do not obey simple charge-balancing or thermodynamic stability criteria, such as meta-

stable materials^{43,44}. Indeed, as illustrated in Fig. 2b, 84% of materials discovered between 1920 and 1930 were charge-balanced, compared to only 38% of materials discovered between 2010 and 2020. In particular, since DFT approaches to material discovery are designed for targeting ground-state structures, there is a growing need to develop more sophisticated computational methodologies that can capture a broad range of synthesizability factors.

To quantify the performance of *SynthNN* at predicting synthesizable materials for future material discovery efforts, we train *SynthNN* on training sets where the positive examples are materials that were synthesized before a given decade and the negative examples are artificially generated as before. We then evaluate the performance of *SynthNN* on a test set that consists of materials that were synthesized in the decades after the materials in the training set. Since these test sets consist of only synthesized materials, we quantify the model performance by calculating the fraction of future synthesized materials that are predicted by our model to be synthesizable (i.e., recall). We evaluate this test recall at a decision threshold that corresponds to a precision that is 5× better than random guessing on a validation set that is drawn from the same distribution as the training data (Fig. 4).

Each line of Fig. 4 illustrates how *SynthNN*'s ability to predict synthesizability changes as more materials are used in model training. As *SynthNN* is exposed to more materials and more diverse chemistries, we observe a considerable improvement in *SynthNN*'s ability to predict materials that will be synthesized in both the current decade and future decades. In the most recent example, training *SynthNN* on all materials synthesized before 2010 results in a model that can correctly identify 80% of all materials that were synthesized between 2010 and 2019. On the basis of the observed trend that model performance increases with the number of materials used in training, we anticipate that the continued discovery of new materials will further improve the ability of *SynthNN* to predict future materials. Nevertheless, it is important to recognize that we observe a notable reduction in recall when predicting the synthesizability of materials in future decades, compared to the current decade. This result suggests that *SynthNN* performs better at identifying synthesizable materials that are similar to the materials seen in training than for materials that are significantly different from any materials that have been previously synthesized.

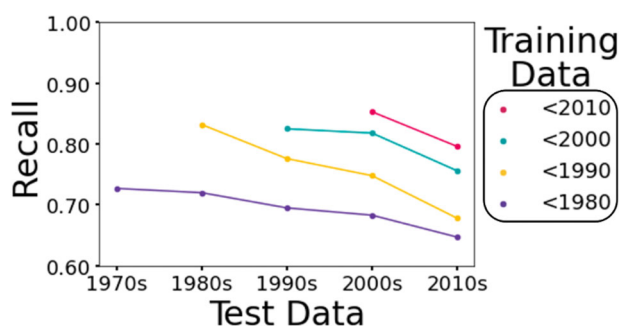


Fig. 4 Performance of *SynthNN* model at discovering future materials. The performance of *SynthNN* for recalling the materials synthesized in a decade, when trained on all the materials synthesized in earlier decades. For example, the '<1980 Training Data' and '1990s Test Data' entry gives the recall of a model trained on all materials synthesized before 1980, when tested on all materials synthesized between 1990 and 1999. The first entry in each row therefore corresponds to the in-distribution test recall, whereas all other entries are the out-of-distribution test recall. In all cases, $N_{\text{synth}} = 20$ is used in training. To allow comparison between models trained on different datasets, the recall values are reported for a precision of 5/21 (5× the random guessing baseline). For each data point, the hyperparameters used are taken from the best-performing model trained on materials from all decades (as shown in Supplementary Table 1).

Model interpretability

Now that we have explored the performance of our model and its potential applications, we now turn our focus to obtaining a better understanding of the inner workings of our model. Although deep neural networks often outperform simpler models (such as random forests or linear models), obtaining interpretable predictions from neural networks is notoriously challenging^{45–48}. Interpretable predictions are particularly important for the task of identifying synthesizable materials since unsuccessful syntheses are extremely costly. Understanding the physically motivated principles that guide the model's decisions can therefore alleviate concerns that the predictions may be the result of an unphysical computational artifact.

Towards this goal, we obtain model interpretability through analyzing both the atomic embeddings that are used to represent the chemical formulas, the hidden layer embeddings of the neural network, and the model outputs. We gain insight into the role of the atomic coefficients by plotting the synthesizability predictions of *SynthNN* for four representative chemical families with varying atomic ratios (Fig. 5a). First, there is a notable distinction between the ionic compounds (Li_xCl_y , Li_xO_y , Fe_xCl_y) and the covalent compound (Li_xGe_y). The Li_xCl_y and Li_xO_y synthesizability predictions exhibit a single sharp peak centered at the compositions that correspond to the most stable oxidation states, indicating the enforcement of *learned* charge-balancing criteria. By comparison, the Li_xGe_y predictions display a much broader range of synthesizable oxidation states, consistent with the covalent nature of this chemical family. Interestingly, the synthesizability predictions of the Fe_xCl_y family not only capture both stable oxidation states (Fe^{3+} and Fe^{2+}), but the higher synthesizability prediction output for FeCl_3 is also consistent with the greater stability of the Fe^{3+} oxidation state⁴⁹. Whereas the results in Fig. 5a are intended to be an instructive example of how charge-balancing influences *SynthNN* synthesizability predictions, we generalize this result (see Supplementary Note 5 and Supplementary Fig. 11) to show that *SynthNN* applies charge-balancing constraints more frequently to ionic materials than non-ionic materials. Taken together, these results suggest that *SynthNN* learns charge-balanced as a rule for predicting synthesizability. However, rather than indiscriminately applying charge-balancing criteria to all predictions, *SynthNN* learns that charge-balancing is most appropriate for ionic compounds. The charge-balancing criteria is relaxed for non-ionic compounds, leading to better generalization across the whole chemical composition space.

Next, we utilize t-distributed stochastic neighbor embedding (t-SNE) to visualize the hidden layer embeddings of the neural network (Fig. 5b). We find that atom types are clustered in a manner that closely resembles traditional periodic table classifications, despite the fact that no periodic table information is provided to our model in training. Whereas the periodic table was originally formulated from an understanding of atomic structure, *SynthNN* is able to derive chemical characteristics of elements through a statistical comparison of the types of formulas that each element can form. Based on these embeddings, we can infer that the predictions of *SynthNN* are partially based on chemical analogy. For example, Fig. 5b shows a clustering of Li_2S with Li_2Se and Li_3S with Li_3Se . This clustering suggests that the synthesizability predictions are partially informed by the synthesizability of chemically analogous materials. A similar clustering of periodic table classifications is observed by visualizing the learned atomic embeddings (Supplementary Fig. 10). This approach bears some similarity to the common material discovery strategy of substituting chemically analogous elements⁵⁰, but we emphasize that *SynthNN* is considerably more flexible since this principle is not strictly enforced for all materials and *SynthNN* is utilizing additional learned criteria for generating synthesizability predictions (Fig. 5a, b).

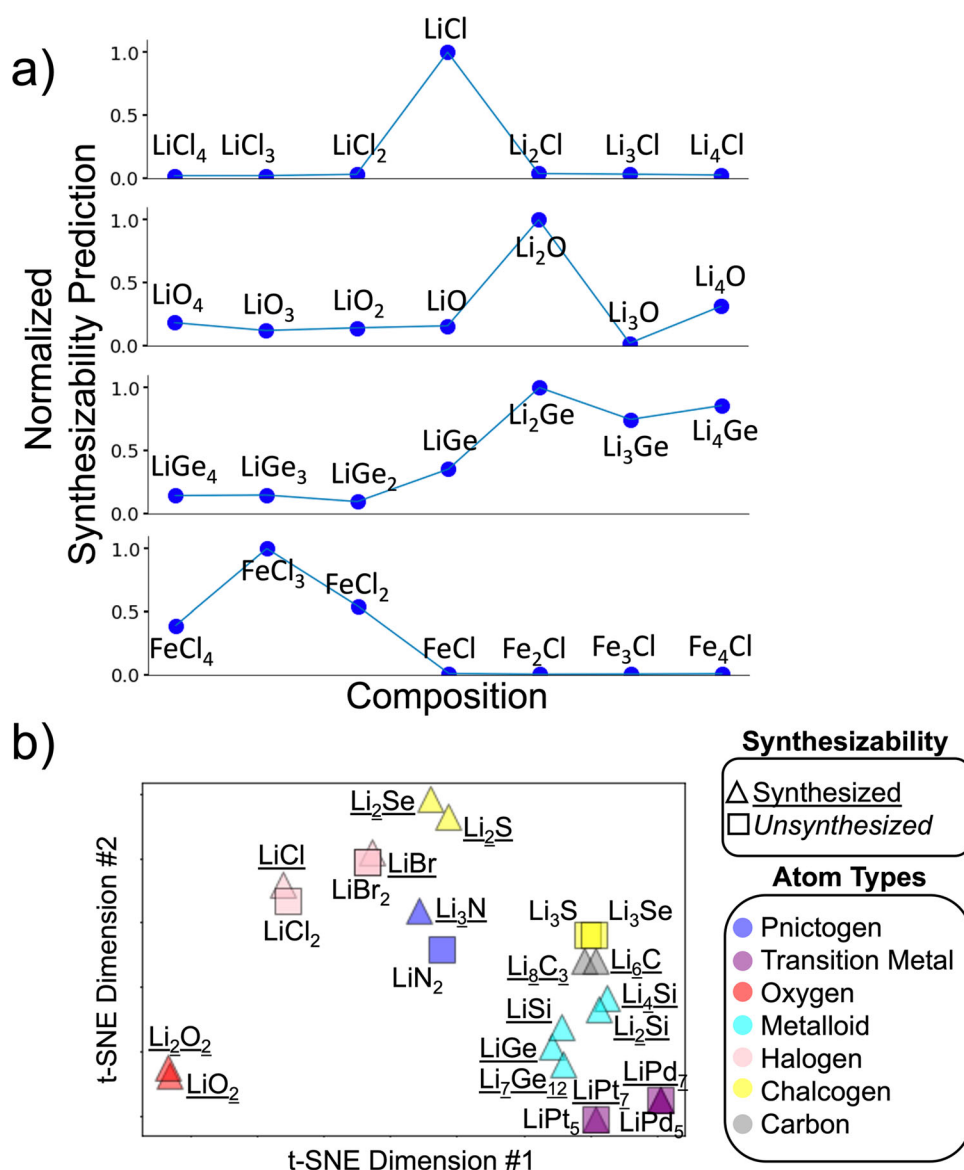


Fig. 5 Extracting interpretability from *SynthNN* predictions. **a** Normalized synthesizability predictions produced by *SynthNN* for a series of chemical compositions from 4 chemical families (Li_xCl_y , Li_xO_y , Li_xGe_y , Fe_xCl_y). All synthesizability predictions are normalized within each chemical family. **b** Two-dimensional representations of the second hidden layer embeddings for representative binary Li chemical compositions obtained with t-SNE.

Although it is not possible to fully capture all of the factors that *SynthNN* utilizes to generate synthesizability predictions, we have shown that the predictions are influenced by the chemical formula's number of unique atom types, thermodynamic stability, charge balance, similarity to chemically analogous formulas and atomic composition (Figs. 3b, 5a, b, and Supplementary Figs. 8–10). Although previous material discovery efforts have employed these metrics, it is notable that the current approach allows for all of these relevant factors to be seamlessly combined in a manner that does not require a manual weighting of these important considerations. Furthermore, criteria such as charge-balancing are flexibly accounted for—only being applied to the materials where it is deemed to be relevant (Fig. 5a).

DISCUSSION

Recent estimates have postulated that approximately 10^{10} – 10^{100} total unique inorganic materials may exist, which is prohibitively

large for discovering materials through an iterative search^{8,51}. Considerable efforts have focused on identifying new materials in this vast search space by performing millions of high-throughput density functional theory (DFT) calculations. Despite considerable advances in the accuracy of DFT calculations, this approach to materials discovery will always be hindered by the extent to which DFT can provide a reliable model of real material systems, as well as the inability of DFT to capture relevant, but incalculable synthesizability criteria. In this work, we have developed an alternative route towards materials discovery that is based on the idea that a meaningful synthesizability classifier can be learned directly from the distribution of previously synthesized chemical compositions. By augmenting this dataset with unsynthesized compositions, we develop a neural network-based synthesizability classifier that utilizes learned compositional features to extract an optimal set of descriptors for the task of predicting synthesizability. In this way, *SynthNN* learns a holistic array of factors that influence synthesizability, beyond what can be captured by DFT.

Our results have shown that this methodology affords significantly more accurate synthesizability predictions than computational approaches based on charge-balancing and DFT calculated formation energies, as well as the predictions of expert human chemists. Importantly, *SynthNN* is also several orders of magnitude faster than predicting synthesizability by calculating formation energies with DFT. We anticipate that *SynthNN* can therefore be used to rapidly search across unexplored regions of chemical space to target fruitful regions of chemical composition space much faster and more accurately than was previously possible. The predictive capabilities of *SynthNN* also have far-reaching implications for realizing an improvement in the reliability and success of computational and experimental materials discovery efforts. Based on the results shown in Fig. 2, utilizing *SynthNN* in computational discovery efforts instead of charge balancing would be expected to yield a 2.6× increase in the precision of identifying synthesizable novel materials. In turn, this is expected to directly translate into a 2.6× reduction in the rate of unsuccessful synthesis attempts, thereby saving years of wasted experimental effort.

METHODS

Synthesizability dataset

The Inorganic Crystal Structure Database (ICSD) is a materials repository that contains inorganic materials that have been both synthesized and structurally characterized. The positive examples that our model is trained on is the set of 53,594 unique binary, ternary, and quaternary compositions in the ICSD database that do not contain fractional stoichiometric coefficients. These positive examples are comprised of 8,194 binary compounds, 26,218 ternary compounds, and 19,182 quaternary compounds. This data was pulled from the ICSD database in October 2020. Due to continuous discovery of new materials, we note that the materials available in the ICSD database are subject to change over time.

Unsynthesized formulas are sampled such that the relative proportion of binary, ternary, and quaternary formulas is the same as in the synthesized materials in our dataset. After the number of unique elements in the unsynthesized formula is selected, the elemental composition is sampled according to the same atomic abundance as the set of positive examples. For example, if 4% of synthesized formulas contain Li, then unsynthesized formulas are generated with a 4% chance of containing Li. Finally, we generate coefficients for each atom in the formula by uniformly sampling between 1 and 20. After generating each formula, we check to ensure the formula, or any multiple of the formula (up to a factor of 20), is not present in the list of synthesized formulas or the set of previously generated unsynthesized formulas. We emphasize that some of these artificially generated materials could be synthesizable, but are absent from the ICSD database or have yet to be synthesized.

Atom2Vec

The methodology for obtaining learned atomic embeddings is derived from the approach developed by Cubuk et al and Zhou et al.^{7,24} Each chemical formula in the Synthesizability Dataset is represented by a normalized (94×1) composition vector, where the rows correspond to relative atomic fraction of each element type. As an example, the composition vector for Na_2O would have 2/3 in the 11th row, 1/3 in the 8th row, and zeroes elsewhere. Then, this composition vector is embedded by performing element-wise multiplication against the ($94 \times M$)-dimensional learned atom embedding matrix, where M is a hyperparameter (Table 1). This yields a ($94 \times M$)-dimensional embedding of the material which is then reduced by averaging across each column of the matrix. This reduced embedding is then used as the input to the deep neural network, described below. The learned atom embedding matrix is randomly initialized and trained alongside all other model parameters.

Neural network model architecture

The *SynthNN* model used in this work is a 3-layer deep neural network originally implemented in TensorFlow 1.12.0. The first two layers utilize hyperbolic tangent (tanh) activation functions, whereas the final layer uses a softmax activation function. The number of hidden units in the first two layers are hyperparameters that are sampled from [30,40,50,60,80]. *SynthNN* is trained with an Adam optimizer⁵² with the learning rate as a hyperparameter that is sampled from [2×10^{-2} , 5×10^{-3} , 2×10^{-3} , 5×10^{-4} , 2×10^{-4}] and a cross-entropy loss function.

We train *SynthNN* on a 90:5:5 split of the Synthesizability Dataset, for a predetermined ratio of synthesized:unsynthesized chemical formulas (N_{synth} in Table 1). To perform semi-supervised learning, we first train *SynthNN* for N_{init} initial iterations by treating the unsynthesized materials as negative examples, where N_{init} is a hyperparameter sampled from [2×10^4 , 4×10^4 , 6×10^4 , 8×10^4 , 1×10^5]. After this initial training stage, we reweight all of the unsynthesized materials according to the procedure specified by Elkan and Noto⁵³. With this approach, all unsynthesized formulas are duplicated to give one positively labeled example and one negatively labeled example. The positive and negative duplicates are then weighted according to the probability that they belong to their respective classes. This approach therefore helps to overcome the incomplete labeling of unsynthesized examples by allowing unsynthesized materials to be treated as positive examples if there is a high probability that they are synthesizable. The model is then trained on this reweighted dataset for an additional 8×10^5 steps (see Supplementary Fig. 6). The final model parameters of each training run are then taken from the step in training that achieves the highest validation accuracy. All hyperparameters are tuned by performing a grid-search and choosing optimal values according to the area under a precision-recall curve (AUC) for a $N_{\text{synth}} = 20$ validation set (see Supplementary Note 1). The hyperparameters used in this model and their range of sampled values are given in Table 1 below. For each value of M and N_{synth} , at least 20 training runs were performed with all other hyperparameters randomly sampled. The training run that achieved the best performance on the validation set for each value of M and N_{synth} is given in Supplementary Tables 1–2.

Model performance evaluations

The model performance evaluations shown in Fig. 2 are all calculated by treating the artificially generated unsynthesized materials as negative examples. Specifically, true positives are synthesized materials predicted to be synthesizable, false positives are unsynthesized materials predicted to be synthesizable, true negatives are unsynthesized materials predicted to be unsynthesizable, and false negatives are synthesized materials predicted to be unsynthesizable. Following from these definitions, the performance evaluations shown throughout this paper take on the definitions used in standard classification tasks.

$$\text{Synthesized Material Precision} = \frac{\text{True positives}}{(\text{True positives} + \text{False positives})} \quad (1)$$

$$\text{Unsynthesized Material Precision} = \frac{\text{True negatives}}{(\text{True negatives} + \text{False negatives})} \quad (2)$$

$$\text{Recall} = \frac{\text{True positives}}{(\text{True positives} + \text{False negatives})} \quad (3)$$

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{(\text{True positives} + \text{False negatives} + \text{True negatives} + \text{False positives})} \quad (4)$$

$$\text{F1 - score} = \frac{2 * (\text{Synthesized Material Precision}) * (\text{Recall})}{(\text{Synthesized Material Precision}) + (\text{Recall})} \quad (5)$$

In the context of predicting the synthesizability of materials, synthesized material precision therefore corresponds to the

fraction of materials predicted to be synthesizable that can be successfully experimentally synthesized. A high precision model therefore minimizes the likelihood that materials predicted to be synthesizable will not be able to be experimentally synthesized in the lab. On the other hand, recall corresponds to the fraction of all synthesizable materials that are successfully predicted to be synthesizable by the model. Achieving a model with high recall is therefore desirable to ensure that a high proportion of all synthesizable materials in the chemical space of interest are captured by the model.

Importantly, the incomplete labeling of the unsynthesized materials results in an overestimation of false positive predictions and an underestimation of true positive predictions since materials that are synthesizable, but have yet to be synthesized will be incorrectly treated as false positives instead of true positives. In terms of these performance metrics, this results in an underestimation of the synthesized material precision and accuracy relative to the true model performance. However, the recall is unaffected since recall only depends on the set of synthesized materials, for which we have correct class labels. Since the F1-score is harmonic mean of precision and recall, it will also be underestimated due to the underestimation of precision. Finally, the area under the curve of a precision-recall curve (i.e., Supplementary Figs. 1–5) will be underestimated due to the underestimation of precision.

DATA AVAILABILITY

The Synthesizability Dataset used during the current study can be obtained from the Inorganic Crystal Structure Database²⁵. The formation energy dataset used for training Roost (Fig. 2c) is included at <https://github.com/antoniuk1/SynthNN>.

CODE AVAILABILITY

Code used for training *SynthNN* and generating synthesizability predictions is available at <https://github.com/antoniuk1/SynthNN>. The code for Roost is available at <https://github.com/CompRPhyS/roost>.

Received: 10 February 2023; Accepted: 8 August 2023;

Published online: 25 August 2023

REFERENCES

- Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *Npj Comput. Mater.* **6**, 1–13 (2020).
- Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Corey, E. J., Cramer, R. D. I. & Howe, W. J. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. *J. Am. Chem. Soc.* **94**, 440–459 (1972).
- Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
- Chamorro, J. R. & McQueen, T. M. Progress toward solid state synthesis by design. *Acc. Chem. Res.* **51**, 2918–2925 (2018).
- Turnbull, D. & Vonnegut, B. Nucleation catalysis. *Ind. Eng. Chem.* **44**, 1292–1298 (1952).
- Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
- Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chemistry* **1**, 617–627 (2016).
- Dan, Y. et al. Generative adversarial networks (GAN) based efficient sampling of chemical space for inverse design of inorganic materials. *Npj Comput. Mater.* **6**, 84 (2020).
- Sun, W. et al. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
- Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-based synthesizability prediction of crystals using partially supervised learning. *J. Am. Chem. Soc.* **142**, 18836–18843 (2020).
- Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
- Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
- Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 1–7 (2016).
- Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *Npj Comput. Mater.* **6**, 138 (2020).
- Jha, D. et al. ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 17593 (2018).
- Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).
- Davariashiyani, A., Kadkhodaie, Z. & Kadkhodaie, S. Predicting synthesizability of crystalline materials via deep learning. *Commun. Mater.* **2**, 1–11 (2021).
- Aykol, M. et al. Network analysis of synthesizable materials discovery. *Nat. Commun.* **10**, 2018 (2019).
- Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
- Kononova, O. et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
- Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
- Kim, E. et al. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
- Zhou, Q. et al. Atom2Vec: learning atoms for materials discovery. *Proc. Natl Acad. Sci. USA* **115**, E6411–E6417 (2018).
- Levin, I. NIST Inorganic Crystal Structure Database (ICSD). (2020) <https://doi.org/10.18434/M32147>.
- Cheon, G. et al. Revealing the spectrum of unknown layered materials with superhuman predictive abilities. *J. Phys. Chem. Lett.* **9**, 6967–6972 (2018).
- Frey, N. C. et al. Prediction of synthesis of 2D metal carbides and nitrides (MXenes) and their precursors with positive and unlabeled machine learning. *ACS Nano* **13**, 3031–3041 (2019).
- Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760 (2020).
- Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *Npj Comput. Mater.* **6**, 1–11 (2020).
- Oganov, A. R., Lyakhov, A. O. & Valle, M. How evolutionary crystal structure prediction works—and why. *Acc. Chem. Res.* **44**, 227–237 (2011).
- Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
- Cheon, G., Yang, L., McCloskey, K., Reed, E. J. & Cubuk, E. D. Crystal Structure Search with Random Relaxations Using Graph Networks. *ArXiv201202920* (2020).
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J. & Greenspan, H. Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification. *ArXiv180102385 Cs* (2018).
- Wang, X., Man, Z., You, M. & Shen, C. Adversarial Generation of Training Examples: Applications to Moving Vehicle License Plate Recognition. *ArXiv170703124 Cs* (2017).
- Marmanis, D. et al. Artificial Generation of Big Data for Improving Image Classification: A Generative Adversarial Network Approach on SAR Data. *ArXiv171102010 Cs* (2017).
- Moore, T. & Clayton, R. Evaluating the Wisdom of Crowds in Assessing Phishing Websites. In *Financial Cryptography and Data Security* (ed. Tsudik, G.) 16–30 (Springer, 2008). https://doi.org/10.1007/978-3-540-85230-8_2.
- Budescu, D. V. & Chen, E. Identifying expertise to extract the wisdom of crowds. *Manag. Sci.* **61**, 267–280 (2015).
- Steyvers, M., Miller, B., Hemmer, P. & Lee, M. The Wisdom of Crowds in the Recollection of Order Information. In *Advances in Neural Information Processing Systems* vol. 22 (Curran Associates, Inc., 2009).
- Hertwig, R. Tapping into the Wisdom of the Crowd—with Confidence. *Science* **336**, 303–304 (2012).
- Kostelnik, T. I. & Orvig, C. Radioactive main group and rare earth metals for imaging and therapy. *Chem. Rev.* **119**, 902–956 (2019).
- Martinez-Gomez, N. C., Vu, H. N. & Skovran, E. Lanthanide chemistry: from coordination in chemical complexes shaping our technology to coordination in enzymes shaping bacterial metabolism. *Inorg. Chem.* **55**, 10083–10089 (2016).
- Zhang, W. et al. Unexpected stable stoichiometries of sodium chlorides. *Science* **342**, 1502–1505 (2013).
- Hong, J. et al. Metastable hexagonal close-packed palladium hydride in liquid cell TEM. *Nature* **603**, 631–636 (2022).

44. Gopalakrishnan, J. Chimie Douce approaches to the synthesis of metastable oxide materials. *Chem. Mater.* **7**, 1265–1275 (1995).
45. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
46. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 1–36 (2019).
47. Zhang, P., Shen, H. & Zhai, H. Machine learning topological invariants with neural networks. *Phys. Rev. Lett.* **120**, 066401 (2018).
48. Xie, T. & Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *J. Chem. Phys.* **149**, 174111 (2018).
49. Tomy, S. et al. Indefinitely stable iron(IV) cage complexes formed in water by air oxidation. *Nat. Commun.* **8**, 14099 (2017).
50. Davies, D. W., Butler, K. T., Isayev, O. & Walsh, A. Materials discovery by chemical analogy: role of oxidation states in structure prediction. *Faraday Discuss.* **211**, 553–568 (2018).
51. Walsh, A. The quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
52. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv1412.6980 Cs* (2014).
53. Elkan, C. & Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* 213–220 (ACM Press, 2008). <https://doi.org/10.1145/1401890.1401920>.
54. Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).

ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. We would like to thank Prof. Tony Heinz for the original project inspiration and the human participants of the Synthesizability Quiz.

AUTHOR CONTRIBUTIONS

The project was conceived by E.R.A., G.C., and E.J.R. E.R.A., G.W., D.B., G.C., and W.C. contributed to the development of the machine learning models. E.R.A. performed

the machine learning experiments and wrote the initial manuscript. All authors contributed to editing the manuscript.

COMPETING INTERESTS

All authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-01114-4>.

Correspondence and requests for materials should be addressed to Evan R. Antoniuk.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023