npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

# Perovskite synthesizability using graph neural networks

Geun Ho Gu[1,2,5], Jidon Jang[1,5], Juhwan Noh[1,5], Aron Walsh[3,4] and Yousung Jung[1] ✉

Perovskite is an important material type in geophysics and for technologically important applications. However, the number of synthetic perovskites remains relatively small. To accelerate the high-throughput discovery of perovskites, we propose a graph neural network model to assess their synthesizability. Our trained model shows a promising 0.957 out-of-sample true positive rate, significantly improving over empirical rule-based methods. Further validation is established by demonstrating that a significant portion of the virtual crystals that are predicted to be synthesizable have already been indeed synthesized in literature, and those with the lowest synthesizability scores have not been reported. While previous empirical strategies are mainly applicable to metal oxides, our model is general and capable of predicting the synthesizability across all classes of perovskites, including chalcogenide, halide, and hydride perovskites, as well as anti-perovskites. We apply the method to identify synthesizable perovskite candidates for two potential applications, the Li-rich ion conductors and metal halide optical materials that can be tested experimentally.

*npj Computational Materials* (2022)8:71 ; https://doi.org/10.1038/s41524-022-00757-z

## INTRODUCTION

The discovery of novel functional materials is a major goal in materials science. Advancement in electronic structure calculations and the development of digital crystal databases have led to the successful discovery of some new functional materials via high-throughput screening (HTS)[1–6]. The HTS is typically conducted in hierarchical stages with increasing accuracy and cost, often starting with the screening of density functional theory (DFT) database of previously synthesized materials, followed by high-level DFT refinements and experimental verifications[2,6]. To expand the scope, databases such as Materials Project[7], OQMD[8], and AFLOW[9] have been collecting a large number of virtual crystals, which are ground-state structures in silico but not yet experimentally synthesized. Some of the promising virtual crystals indeed have been synthesized[10–12], demonstrating the validity of a virtual screening strategy to discover new materials.

Many, if not the most, screened virtual materials are not experimentally realized[11], thus assessing synthesizability has been an important subject[13–20]. Typically, the synthesizability of virtual materials is assessed using the energy above convex hull[6,11,21,22]. As well recognized, however, the latter thermodynamic metric is insufficient for assessing synthesizability as the synthesis kinetics and growth conditions are largely neglected in that approach (e.g., selection of precursors, annealing temperature and duration, external pressure, and so on)[11]. Therefore, developing a generalized and more reliable method to predict the synthesizability of a candidate crystal can significantly accelerate the high-throughput discovery of new materials.

A binary classification (positive and negative labeling) may be used to predict stability. However, such positive and negative learning cannot be used to predict the synthesizability as there is no negative ("unsynthesizable") crystal data, since the inability to synthesize a hypothetical crystal is difficult to know a priori. Hence, databases only have previously synthesized crystals (positive) and potentially synthesizable crystals (unlabeled). The positive-unlabeled (PU) semi-supervised classification methods aim to predict positivity for problems where the negative data are hard to obtain[23–25]. Indeed, a transductive PU-learning method[26] has been used recently[27,28] to predict the synthesizability score, called crystal-likeness (CL) score, of unlabeled virtual crystal structures in the MP database. The model showed a respectable out-of-sample positive data prediction accuracy of around 87%[27]. Since the method uses crystal graph convolutions to encode material information, it can be seen as a structure-based synthesizability prediction model, in comparison to conventional thermodynamics-based estimations.

While the previous work demonstrated the proof-of-concept for synthesizability prediction, the model accuracy for specific subdomains of chemical space such as perovskites (74%) was below the overall accuracy (87%)[29,30]. Since perovskites are increasingly receiving wide attention for their applications in photovoltaics[31–33], light-emitting diode[34–36], magnetic materials[37,38], superconductors[39–41], and Li-ion conductors[42], developing a perovskite focused model with improved accuracy would be invaluable for more efficient materials discovery[10,43].

Indeed, several previous synthesis models have focused on perovskites. Heuristic-based Goldschmidt tolerance factor is commonly implemented to predict stability for ionic perovskite[44] based on the ionic radii of the constituent elements. Similarly, Bartel et al. developed a machine learning (SISSO)-determined tolerance factor to classify structure type (perovskite vs. non-perovskite) for the ionic perovskites[45]. In addition, gradient boosting decision tree[46–48], support vector machine[49], random forest classification[47,50], and combination of multiple models[51] were used to develop similar classification models. However, previous models focused largely on metal oxide perovskites and mostly relied on the Shannon ionic radii database[52], making the consideration of the perovskites with more covalent bonding[53] or the anti-perovskites[54] difficult due to the limited scope of Shannon's table[52]. Potentially, training a generalized deep learning model could address these deficiencies.

The previous study has shown that training the model with a particular domain of materials can improve the model accuracy[55]. Such domain-specific learning could also improve the synthesizability prediction for perovskites as well. Another challenge in

[1]Department of Chemical and Biomolecular Engineering (BK21 four), KAIST, 291 Daehak-ro 34141, Daejeon 305-335, South Korea. [2]School of Energy Technology, Korea Institute of Energy Technology, 200 Hyuksin-ro, Naju 58330, South Korea. [3]Department of Materials, Imperial College London, London SW7 2AZ, UK. [4]Department of Materials Science and Engineering, Yonsei University, Seoul 03722, South Korea. [5]These authors contributed equally: Geun Ho Gu, Jidon Jang, and Juhwan Noh. ✉email: ysjn@kaist.ac.kr

applying the PU learning framework[27] to a chosen structure type is the small data size for the prototype. Transfer learning[56] is a widely-used strategy to train a deep neural network with a small data set, where a more general model is first developed with a large data set that contains the target domain, and the knowledge of this general model is transferred to a new model and the portion of the model is retrained with the target domain data set[57].

Here, we combine positive-unlabeled learning[26], domain-specific learning[55], and transfer learning[56] to develop the synthesizability prediction model of perovskites with a high practical accuracy. We pre-train the graph neural network with the Materials Project database and retrain the portion of the model with the smaller perovskites dataset. 943 previously synthesized perovskite crystals and 11,964 virtual perovskites collected from Materials Project (MP), OQMD, and AFLOW databases were used for learning. Our model shows a high out-of-sample positive data accuracy of 95.7%, compared to those of the non-domain specific original model around 74.0%. Our model predicted 962 materials out of 11,964 virtual perovskites as synthesizable, and 179 virtual crystals of those have indeed been synthesized in literature. Compared to the previous ionic perovskite-focused models, our model is capable of predicting the synthesizability of all types of perovskites in the dataset, including anti-perovskites where the anion and cation occupation is inverted. We furthermore suggest promising Li-rich anti-perovskites and metal halides as candidates for solid-state electrolyte and photoactive materials discovery, respectively.

## RESULT AND DISCUSSION

### Development of PU learning model

The inorganic crystal data from the MP[7] database, retrieved in October 2020, consisted of 46,546 crystals with inorganic crystal structure databases (ICSD) id and 79,789 crystals without ICSD id. We considered the 46,546 crystals with the ICSD id and experimental tag synthesizable, and the remaining 79,789 crystals without ICSD id "virtual", as undetermined. These MP data are used to pre-train the model. We then retrieved the perovskite crystals from MP[7], OQMD[8], and AFLOW databases[9] in October 2020 (Fig. 1a). We used the StructureMatcher function in pymatgen[58] and perovskite prototype structures in the AFLOW database[9] to identify and remove duplicate crystals, resulting in 943 synthesized and 11,964 virtual perovskite crystals. The perovskite data are used to train the transferred model.

Both the pre-training and transfer learning are performed using inductive PU learning[26]. To test our model, 10% of randomly sampled synthesized crystals are set aside from both the MP data used for pre-training and the perovskite data used for transfer learning. Thus, we ensure that the test data is not observed for the pre-training stage. With the rest of the data set, we perform the PU

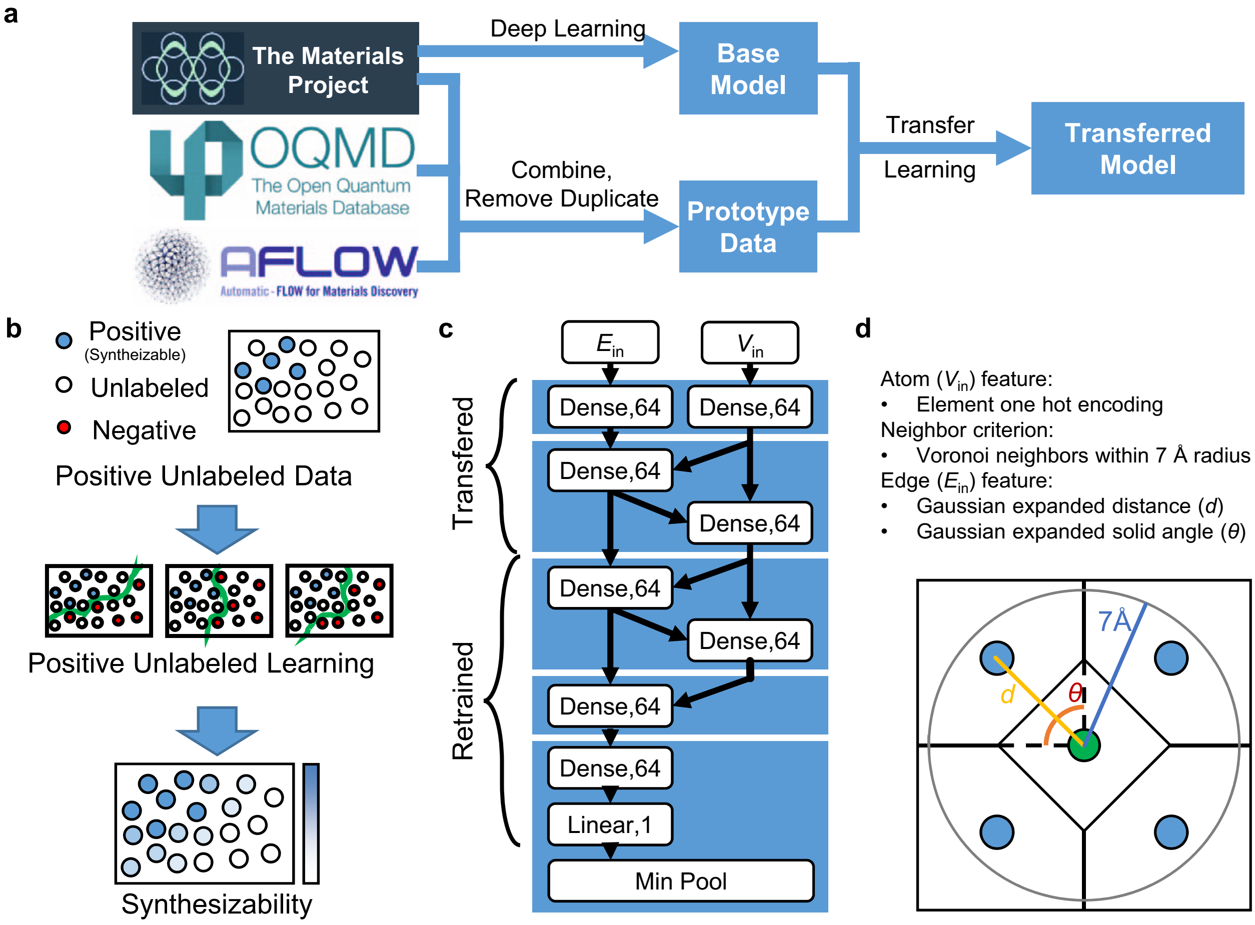

**Fig. 1 Overview of the model development.** **a** Domain-specific transfer learning workflow. The model is first trained with the Materials Project database, and the model is re-trained with the perovskite-only data extracted from the three databases. **b** Positive and unlabeled learning procedure overview. **c** The graph neural network architecture. $E_{in}$ and $V_{in}$ are the atom and edge features. Dense indicates the linear multiplication followed by the softplus activation layer and Linear indicates linear multiplication. The number next to the operation indicates the output feature dimension. Min Pool indicates minimum pooling followed by sigmoid activation. More detail is in the "Methods" section. **d** The crystal representation. Atoms and edges are converted to mathematical representation via featurization.

learning procedure. Here, 10% of the synthesized crystals are randomly sampled and the same number of virtual crystals are randomly sampled, both for the model validation. The rest of the synthesized crystals are used for training, and the same number of virtual crystals are randomly sampled and treated as negative data for the training. This process is repeated 100 times, resulting in an ensemble of 100 models. The key procedure is that, for each model, the training and validation set for the virtual crystals change, whereas those for the synthesized crystals remain fixed. The synthesizability score, which we call CL score[27], is calculated by averaging over the predictions of 100 models. Varying the virtual data set aids in forming the averaged decision boundary as shown conceptually in Fig. 1b.

For the prediction model, we constructed a graph convolutional neural network (GCNN) inspired by MEGNet[59] as shown in Fig. 1c, the detail of which is provided in the method section including the crystal featurization. Our model calculates the CL score between 0 and 1, where the crystals with a high CL score indicate high synthesizability. For practical screening, crystal candidates can be tested in the decreasing order of CL score for the best chance of success. In this work specifically, we empirically set the CL score of 0.5 to calculate metrics such as true positive rate (TPR; true positive/(true positive + false negative)) and also to consider a crystal as a synthesizable candidate. To perform transfer learning, we first pre-train our model with the Materials Project data. Then, the model weights in the encoding layer and the first graphical convolution layer are fixed and the rest of the model is re-trained using the combined perovskite data.

### Model accuracy and validation based on previous experiments

We assess the model by the true positive rate using the held-out positive test set as shown in Fig. 2a. We focus on the TPR since negative data (unsynthesizable) are unavailable. Compared to the MP-trained general synthesizability prediction model, the domain-specific transfer PU learning has significantly higher TPR for perovskites, increased from 0.740 (GCNN + PUL in Fig. 2a) to 0.957 (GCNN + PUL + DSL + TL in Fig. 2a). For comparison, we also tested the CGCNN model in our previous work[27], and found that TPR is 0.595 and 0.957 for the general model and the domain-specific transfer PU learning, respectively, suggesting that the domain-specific transfer learning is more important than the model architecture. We plotted the CL score distribution for the virtual and synthesized crystals (Fig. 2b) to assess perovskite chemical space. The scores for virtual crystals are skewed towards the CL score of 0, and only 962 (1121 considering structure distortion) out of 11,964 virtual perovskites are predicted synthesizable. We find that domain-specific transfer learning can improve the accuracy for oxide-focused chemical space (from 0.837 to 0.930). We note that while TPR can be artificially increased by lowering the threshold probability or developing a naïve model that predicts all crystals synthesizable, such is not the case for our model as 84% of the perovskite crystals are predicted unsynthesizable. Figure 2b demonstrates the CL score distribution for all data and the out-of-sample test data, which also shows that virtual crystals are generally predicted unsynthesizable.

To understand the motive of the model's success, we test the binary classification model where GCNN is trained using a dataset where all unlabeled data is labeled negative, and positive data are oversampled to balance the number of positive and negative data. Here, we find that TPR decreases to 0.361 (GCNN + BC in Fig. 2a) and 0.691 (GCNN + BC + DSL + TL in Fig. 2a) for the MP-trained general model and transfer learning model, respectively. This could be due to the positive data in unlabeled data that are mislabeled negative, thus the data splitting method in PU learning is critical. We also trained a PU-learning model without pre-training with MP data (i.e., without transfer learning), and find that the TPR decreases slightly to 0.947 (GCNN + PUL + DSL in Fig. 2a). Thus, the model success is largely attributed to the domain-specific data set, and the transfer learning scheme contributes marginally for TPR. For the rest of the discussion, we will use results obtained from the best model, GCNN + PUL + DSL + TL.

We further investigated the correlation between the predicted CL score and the energy above hull for all the virtual crystals obtained from each data source as shown with the histogram and violin in Supplementary Fig. 4. The overall data distribution between the CL score and the energy above hull shows a negative correlation (Pearson correlation coefficient of −0.3739). Thus, our model learns the energetic stability (energy above hull) to some extent without explicitly learning these metrics. Interestingly, Supplementary Fig. 4 shows that a significant number of energetically stable perovskites (energy above hull < 0.1 eV/atom) have low CL scores, indicating the difference between the machine-learned synthesizability and the conventional energetic synthesizability metrics.

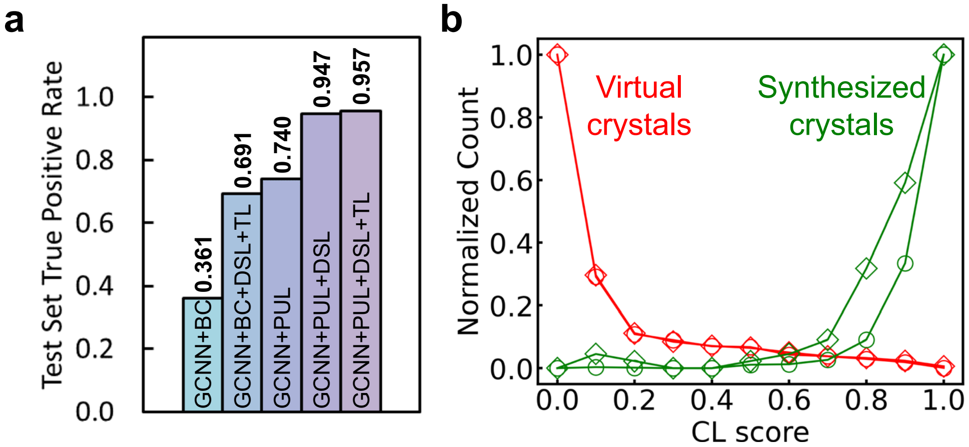

**Fig. 2 Model accuracy and data distribution. a** The out-of-sample true positive rate for perovskites for the various tested models. GCNN indicates graph convolutional network, BC indicates binary classification, PUL indicates positive-unlabeled learning, DSL indicates domain-specific learning, and TL indicates transfer learning. True positive rate is assessed as the performance measure, as the positive data (synthesized crystals) are known, while the negative data (unsynthesizable crystals) are not known. As the unlabeled data (virtual crystals) are available in the database, positive unlabeled learning is implemented to assess synthesizability. **b** The score distribution for the synthesized and virtual crystals. Diamond and circle marks indicate the out-of-sample test data, and all data, respectively. The count is normalized by the highest peak value.

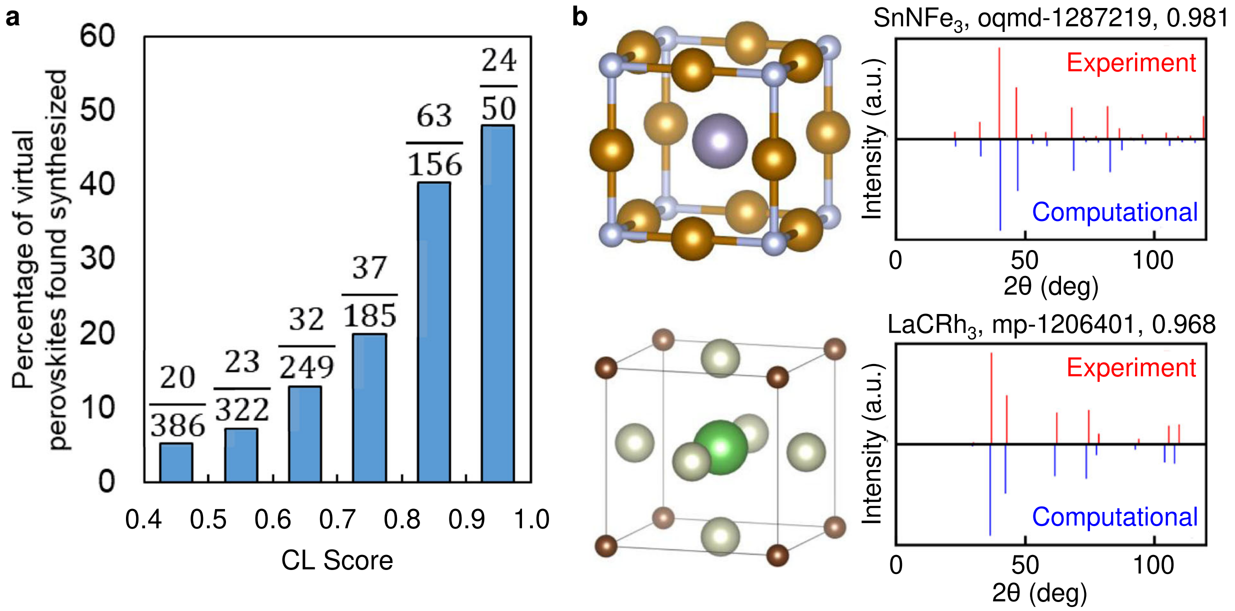

**Fig. 3 Model validation. a** The percentage of virtual perovskites that are found synthesized in the literature. The ratio indicates the number of found over the number of virtual crystals in the range. **b** The structure of virtual crystals and XRD comparison between the experimental and virtual crystals for the top two perovskites previously reported in refs. [71,72]. The full list of virtual crystals and XRD pattern comparison are shown in Supplementary Table 3 and Supplementary Fig. 5.

To further validate our model in practice, we searched the literature for the reported cases of synthesis for the virtual crystals that are predicted synthesizable. We used the XRD patterns to match the synthesized and virtual crystals as shown in Supplementary Fig. 5. We found that 179 out of 962 synthesizable virtual perovskite compounds have been synthesized before (Supplementary Fig. 5 and Supplementary Table 3). In further analysis, the percentage plot of the found virtual crystals by the CL score in Fig. 3a shows an interesting trend that the ratio of the previous synthesis increases with the predicted CL scores. Figure 3b shows the two previously synthesized virtual perovskites with the highest synthesizability scores and their respective XRD patterns. We also searched the literature for the 1000 virtual crystals with the lowest CL scores but were not able to find any previous report of their synthesis. To furthermore assess the model's performance for the crystals with an indecisive score, we searched the literature for the crystals with CL scores between 0.4 and 0.5. We found only 20 crystals previously synthesized out of 386 virtual crystals for these crystals, indicating the value of CL score in the indecisive region. While these assessments provide validation for our model, we cannot guarantee the model's high precision (true positive/(true positive + false positive)), as it is difficult to show that our positive predictions are incorrect.

### Comparison with tolerance factor-based models

We compare our models' out-of-sample TPR with the two empirical perovskite discovery strategies, i.e., Goldschmidt rule-based and SISSO-based screening, by the assumption that the materials are considered to be synthesizable if they remain after applying the screening filters. Davies et al.[60] used the Goldschmidt tolerance factor[44]-based screening by assessing the ionic radius of the Shannon table[52]. This screening focused on standard ionic perovskites, where the element of the C site in the ABC$_3$ formula was limited to 7 anions. Since our data contains non-classical ionic perovskites, only 388 out of 943 synthesized perovskites were found to be within their screening scope. For those 388 perovskites that are directly relevant to the Davies et al.'s[60] procedure (see SI), a TPR of 0.863 is obtained using Davies et al's method. Bartel et al.[45] developed and used a SISSO determined tolerance factor that uses the oxidation state and the ionic radius[52]. Only 310 crystals out of 943 perovskites were within their selection of elements, but by reproducing their procedure (see SI), we calculated the TPR of 0.806. Note that the reported[45] TPR (0.936) is different which may be due to the difference in the dataset. Nonetheless, our model's out-of-sample TPR (0.957) is significantly higher (0.806–0.863) than the previous methods for the experimentally synthesized perovskites considered.

Also, our model chooses less synthesizable candidates than the previous strategies. Supplementary Figure 1 compares the Goldschmidt rule-based screening results and CL score and demonstrates that a large portion of the virtual crystals that passes the screening have low CL score. More precisely, our model predicts that 9.4% of the virtual crystals are synthesizable, whereas Davies et al.[60] and Bartel et al.[45] predict that 24.5 and 25.7% are accessible, respectively. Figure 4 compares the 2D elementary map for ABO$_3$ perovskite oxide based on our model and Goldschmidt-based screening of Davies et al.[60] which also shows that fewer candidates are predicted synthesizable by our model. Here, the red and blue boxes indicate the virtual crystals that have been synthesized. This result does not necessarily indicate that our model is more selective, as the synthesizability of crystals is difficult to measure. On the other hand, our model predicts a probability, thus the best candidates can be prioritized.

While the previous strategies focused on the Shannon table[52]-based classical ionic perovskite predictions, our model can predict synthesizability for various perovskite types. We expanded the screening scope of Davies et al.[60] by including the non-anion elements for the C site as well. In this case, the TPR of the screening method is low (0.389) which is attributed to increased covalency in bonds for some elemental combinations[53] and the lack of relevant elemental data in the Shannon tables for anti-perovskites[54]. In addition to classical ionic perovskites, we found unconventional combinations of elements within the 179 virtual perovskites that were found synthesized, types of which are "covalent" perovskites that contain two or more anions (e.g., CsIO$_3$, ClOLi$_3$) with higher covalency in bonds, hydride perovskites that contain hydrogen (e.g., CaCsH$_3$), and anti-perovskites that contain anion in B sites instead of C sites in ABC$_3$ combination

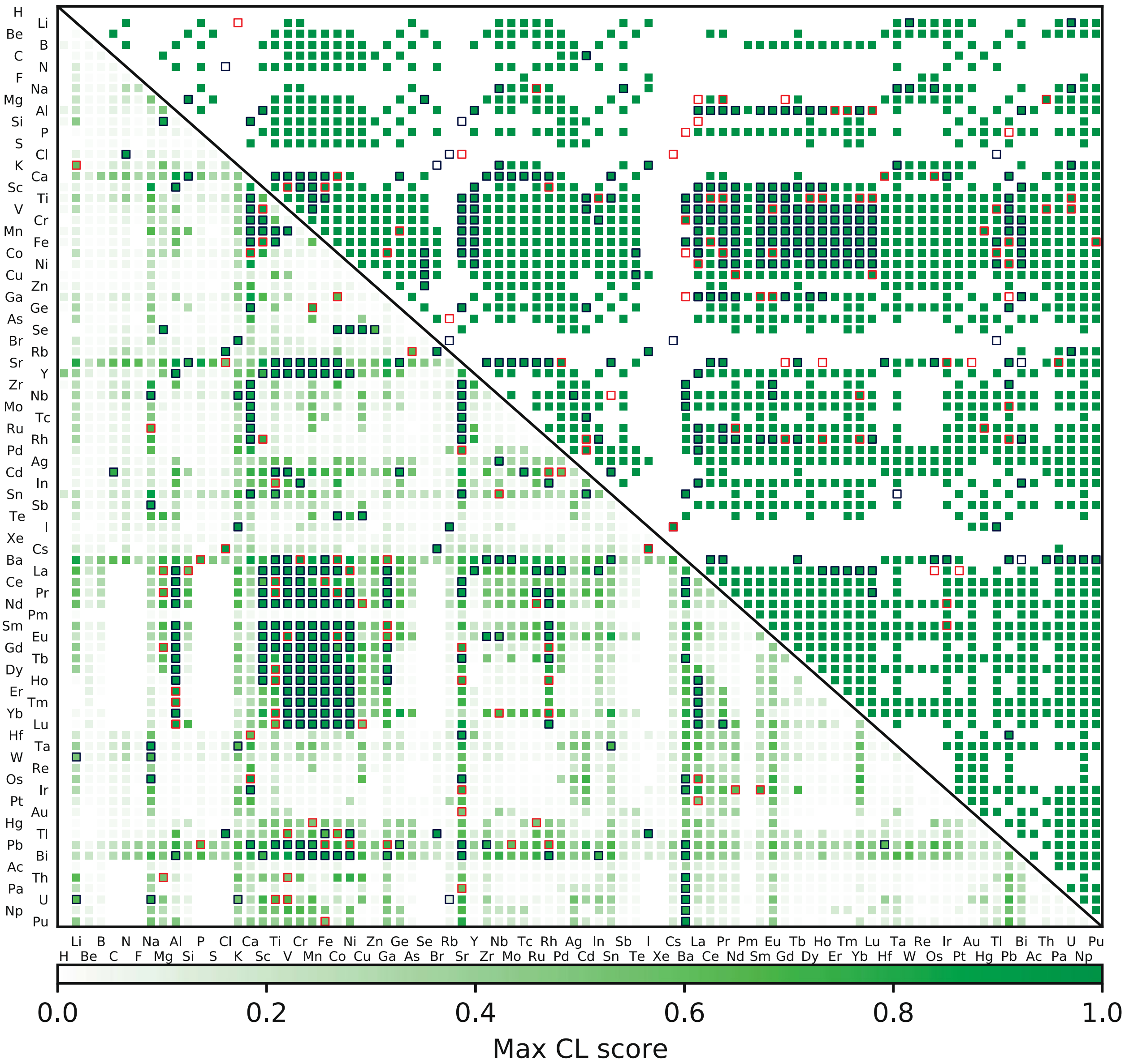

**Fig. 4 Synthesizability of the $ABO_3$ perovskite compounds for our model (lower left triangle) and Goldschmidt-rule-based screening (upper right triangle).** The green color in the lower left triangle indicates the maximum CL score for the perovskites structures from the databases with the given compositions, the color bar of which at the bottom of the figure, and the green color in the upper right triangle indicates that the combination passes the screening. The blue box indicates that the combination has been synthesized before. The red boxes indicate the virtual crystals that were found synthesized previously.

(e.g., $SnNFe_3$) (see Fig. 5). The prediction for these three types of combinations is a new capability of our model that the previous models were not capable of. Figure 5c, d shows the SISSO-based model[45] and Goldschmidt rule-based screening[60] result for the discovered 179 virtual crystals, where we observe that the significant portion is outside-of-scope. Also, Fig. 5b shows that the non-domain-specific model only predicts 101 crystals stable out of the found 179 virtual crystals, showing the value of the domain-specific learning.

## Applications

While perovskite has been studied extensively, Fig. 4a shows that there remain many synthesizable elemental combinations yet to be discovered. We plot the periodic table representation of the synthesizability in Supplementary Fig. 2. Here, the ratios of virtual candidates with CL scores above 0.5 are shown with the given element in the given site. Compared to the classical ionic perovskites, anti-perovskites have high CL scores which contain C, N, O, P in the B site, and a transition metal on the C site. Indeed, we found that a significant number of virtual anti-perovskites have been previously synthesized (Fig. 5), suggesting there may be more opportunities to discover anti-perovskites. Anti-perovskites have shown many interesting properties such as superconductivity[39,40] and magnetism[37,38,61]. Our model suggests that 327 virtual anti-perovskites are synthesizable, which are listed in Supplementary Table 3.

We also selected the synthesizable candidates for two technologically important applications. Metal halide perovskites, namely, $CsPbI_3$, $RbPbI_3$, and $MAPbI_3$ ($MA = CH_3NH_3^+$) have shown many promising applications in photovoltaics and light-emitting diodes in the past decade[34–36]. However, these materials often contain toxic Pb. The semiconducting properties of these perovskites are largely due to the diffuse valence p-orbitals of

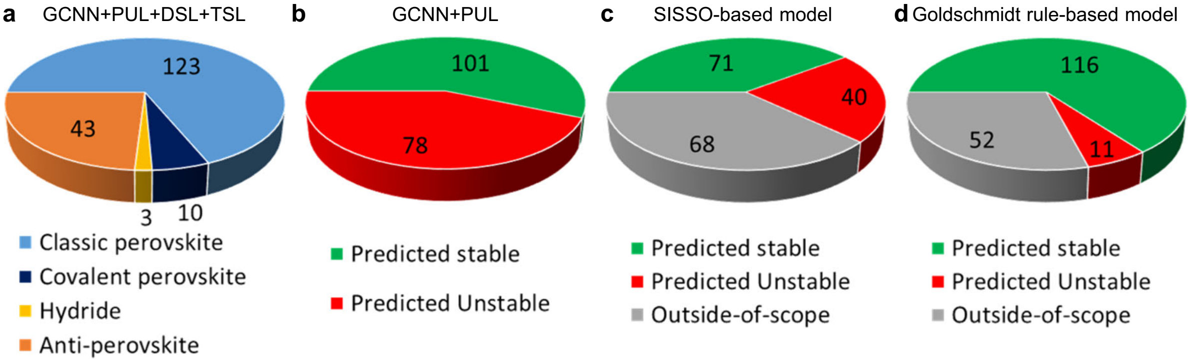

**Fig. 5 Predictions of the other methods for the reported virtual crystals.** **a** The distribution of perovskite types for the 179 virtual perovskites found synthesized. The stability prediction of the 179 compounds using the **b** non-domain-specific MP-trained general model (GCNN+PUL in Fig. 2a), **c** SISSO-based model[45], and **d** Goldschmidt rule-based screening[60]. The $ABC_3$ perovskites were classified based on the following criteria: classical perovskites contain cation in A and B site and anion in C site (e.g., $SrTiO_3$), anti-perovskite contains anion in B site and cation in A and C site (e.g., $SnNFe_3$), covalent perovskite contains two or more anions (e.g., $CsIO_3$, $ClOLi_3$), and hydride contains hydrogen on the C site (e.g., $CaCsH_3$).

**Table 1.** Synthesizable halide perovskites with calculated bandgap in-between 0.7 and 2.0 eV for photovoltaics applications.

| $ABC_3$ | Bandgap [eV] | CL score | Energy above hull (eV/atom) |
|---|---|---|---|
| $RbCoF_3$ | 0.71 | 0.861 | 0.016 |
| $KGeBr_3$ | 0.85 | 0.559 | 0.000 |
| $NPF_3$ | 0.85 | 0.545 | 1.706 |
| $CsNaF_3$ | 0.94 | 0.820 | 0.025 |
| $NaCuCl_3$ | 1.36 | 0.797 | 0.002 |
| $RbCF_3$ | 1.39 | 0.631 | 1.154 |
| $KGeCl_3$ | 1.43 | 0.544 | 0.015 |
| $TlGeCl_3$ | 1.46 | 0.572 | 0.019 |
| $CsPdF_3$ | 1.52 | 0.814 | 0.102 |
| $RbCrF_3$ | 1.82 | 0.560 | 0.097 |
| $RbOF_3$ | 1.92 | 0.586 | 0.374 |
| $InGeCl_3$ | 1.99 | 0.617 | 0.039 |

the halide[62], thus we expect that there are more semiconducting halide perovskites that can be accessed. Our model predicts that 98 virtual metal halides are synthesizable. We further screen these materials by band-gap, using a two-step DFT procedure (PBEsol relaxation followed by HSE06 single point calculation). We found that 43 materials have band gaps as listed in Supplementary Table 1. Particularly, 12 candidates have a bandgap between 0.7 and 2.0 eV, which could be promising for photovoltaics as shown in Table 1 including CL score and the energy above hull. Herein, the majority of the predicted materials (8 of 12 candidates) are thermodynamically stable (energy above hull < 0.1 eV/atom). In addition, as shown in Supplementary Fig. 3b, CL score values of all the predicted materials in Table 2 are overlapped with the CL score distribution of positive data. We note that two materials ($NPF_3$, and $RbCF_3$) are highly unstable (energy above hull > 1.0 eV/atom). While our model has a relatively high true positive rate, the model could make false positive predictions (low precision) as discussed above, resulting in this disparity. We note that many of these compositions contain non-standard chemistries (e.g., $CsNaF_3$ or $RbOF_3$) that would not be identified based on simple electron counting considerations.

Zhao et al.[41] discovered that Li-rich anti-perovskite, $Li_3OCl$ have superionic conductivity for the application of solid battery electrolytes. The high conductivity was achieved due to high Li concentration and the streamlined C-site diffusion pathway, thus the conductivity is expected to be transferable to other Li-rich anti-perovskite such as $Li_3OBr$[63]. We listed 8 Li-rich anti-perovskites with CL score > 0.5 in Supplementary Table 2 including CL score and the energy above hull. While the previously reported $Li_3OBr$ and $Li_3OCl$ are thermodynamically stable (0.012 eV/atom for $Li_3OBr$ and 0.006 eV/atom for $Li_3OCl$), the newly predicted materials in Supplementary Table 2 show low thermodynamic stability (>0.3 eV/atom). Also, a similar disparity is observed for the CL score distribution as well (see Supplementary Fig. 3a), indicating potential difficulties in synthesizing these materials thermodynamically despite being more synthesizable based on the CL scores. This suggests an interesting possibility that the combined use of CL scores and thermodynamic metrics can complement the limitations of each approach and yield more reliable synthesizability predictions.

To summarize, perovskites represent a unique class of materials with desirable physical properties. We have implemented domain-specific transfer PU learning to assess the synthesizability of perovskite materials. Our model demonstrated a 0.957 out-of-sample true positive rate, significantly improving over the previous methods based on geometric factors (0.806–0.863)[45,60]. We searched the literature for the 962 virtual crystals that are predicted synthesizable and found that 179 virtual crystals have been synthesized, adding to the synthesized perovskite pool of 943 crystals in three open crystal databases. The same literature search for the 1000 virtual crystals with the lowest synthesizability scores yielded no synthesized cases, further validating our model. Compared to empirical models based on ionic radii that are most applicable to classical ionic perovskites, our model demonstrates a general ability to assess the synthesizability across all prototypes of perovskites, including the anti-perovskites, covalent perovskites, halides, and hydrides. To this end, we listed promising synthesizable candidates that can expand the materials portfolio for two important applications, i.e., Li-rich ion conductors and metal halide optical materials, which can be tested experimentally. We expect that the proposed domain-specific transfer PU learning would be fruitful to explore the target-specified crystal space for other crystal families and application domains.

## METHODS

### Model architecture and training

The overall architecture of the convolutional neural network is shown in Fig. 1c. $V_{in}$ and $E_{in}$ are the atom and edge/interaction input features to the model. The graph structure of crystals is constructed by assigning edges to Voronoi neighbors within the 7 Å radius of each atom. The atom features are constructed by the one-hot encoding method categorized by the element, and edge features are constructed by Gaussian expansion of distance and Voronoi solid angles as shown in Fig. 1d. These features are encoded with linear multiplication and a softplus activation. The graph convolutional layer contains neighbor edge and atom pooling to make new hidden features. In detail, the new edge features of edge $i$, $E_{\mathrm{out},i}$ are generated by

$$E_{\mathrm{out},i} = \sigma\left(W \cdot \phi\left(V_{\mathrm{in},j}, V_{\mathrm{in},k}, E_{\mathrm{in},i}\right) + \beta\right) \tag{1}$$

where $\sigma$ is the softplus function, $W$ is the linear multiplication weight, $\beta$ is the bias, $\phi$ is the concatenation operator, $j$, and $k$ are the two atoms connecting the edge. The new atom features for atom $i$ are generated by

$$V_{\mathrm{out},i} = \sigma\left(W \cdot \phi\left(V_{\mathrm{in},i}, \sum_{j}^{n_{\mathrm{neighbor}}} \frac{E_{\mathrm{in},j}}{n_{\mathrm{neighbor}}}\right) + \beta\right) \tag{2}$$

where $j$ is the index of edges that are connected to atom $i$. Here, the edge features are averaged and concatenated. The box with "Dense, 64" with two input arrows in Fig. 1c indicates the two convolution operators discussed above. The 64 indicates that the output feature size is 64. The "Dense, 64" with one input arrow indicates a simple activation layer for the feature, $F$,

$$F_{\mathrm{out}} = \sigma\left(W \cdot F_{\mathrm{in}} + \beta\right) \tag{3}$$

For the box with "Linear,1", linear multiplication is used,

$$F_{\mathrm{out}} = W \cdot F_{\mathrm{in}} + \beta \tag{4}$$

resulting in a single element value. The "Min Pool" indicates the minimum pooling operation followed by the sigmoid operation. As discussed above, the intermittent atom and edge features are kept at the element size of 64. We used binary cross-entropy loss function with Adam optimizer[64] to train our model with a batch size of 512. The model is trained to 50 epochs, and the model with the lowest validation loss is selected.

### Bandgap and energy above hull calculations

For all DFT calculations, we performed spin-polarized PBEsol[65,66] calculations with PAW-PBE pseudopotentials[67] as implemented in the plane-wave-based ab initio package, VASP[68]. We selected the PAW potentials as recommended in the MP database[7]. Atomic positions and unit cell parameters are fully relaxed using the conjugate gradient descent method with the convergence criteria of 1.0e−5 eV for the energy and 0.05 eV/Å for the force with 500 eV cut-off energy. Brillouin zone is used with the k-point densities of 1000 k-points per atom using Pymatgen[58]. For the calculations of bandgap using the relaxed structure, we performed HSE06[69] hybrid DFT functional implemented in VASP[68] with a mixing parameter of 0.2. For computational efficiency, we used cut-off energy of 400 eV, and also used a uniform reduction factor for the $q$-point grid of the exact exchange potential is applied (NKRED = 2) with gamma centered even number k-points (with a k-point density of 1000 k-points per atom). For Brillouin zone integration[70], we used Blöchl correction-included tetrahedron method. To calculate energy above hull, we extracted all relevant species in the convex hull diagram from the materials project, and performed PBEsol calculations. The energy above hull is obtained by using the calculated energetics and Pymatgen[58].

## DATA AVAILABILITY

The data used in this paper are publicly available at https://doi.org/10.5281/zenodo.6348980.

## CODE AVAILABILITY

The neural network model used in this paper is publicly available at https://doi.org/10.5281/zenodo.6348980.

Received: 25 November 2021; Accepted: 21 March 2022;
Published online: 20 April 2022

## REFERENCES

1. Zhuo, Y. et al. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nat. Commun.* **9**, 4377 (2018).
2. Newhouse, P. F. et al. Discovery and characterization of a Pourbaix-stable, 1.8 eV direct gap bismuth manganate photoanode. *Chem. Mater.* **29**, 10027–10036 (2017).
3. Muy, S. et al. High-throughput screening of solid-state Li-ion conductors using lattice-dynamics descriptors. *iScience* **16**, 270–282 (2019).
4. Zhou, J. et al. Discovery of hidden classes of layered electrides by extensive high-throughput material screening. *Chem. Mater.* **31**, 1860–1868 (2019).
5. Wei, J. et al. High-throughput screening and classification of layered di-metal chalcogenides. *Nanoscale* **11**, 13924–13933 (2019).
6. Yan, Q. et al. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Natl Acad. Sci. USA* **114**, 3040–3043 (2017).
7. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
8. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *Npj Comput. Mater.* **1**, 15010 (2015).
9. Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
10. Zhu, H. et al. Computational and experimental investigation of TmAgTe2 and XYZ2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *J. Mater. Chem. C* **3**, 10554–10565 (2015).
11. Noh, J. et al. Unveiling new stable manganese-based photoanode materials via theoretical high-throughput screening and experiments. *ChemComm* **55**, 13418–13421 (2019).
12. Bauers, S. R. et al. Ternary nitride semiconductors in the rocksalt crystal structure. *Proc. Natl Acad. Sci. USA* **116**, 14829 (2019).
13. Sun, W. et al. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
14. Aykol, M., Dwaraknath Shyam, S., Sun, W. & Persson Kristin, A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
15. Aykol, M. et al. Network analysis of synthesizable materials discovery. *Nat. Commun.* **10**, 2018 (2019).
16. Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
17. Davariashtiyani, A., Kadkhodaie, Z. & Kadkhodaei, S. Predicting synthesizability of crystalline materials via deep learning. *Commun. Mater.* **2**, 115 (2021).
18. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
19. Jensen, Z. et al. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
20. Tang, B. et al. Machine learning-guided synthesis of advanced inorganic materials. *Mater. Today* **41**, 72–80 (2020).
21. Sendek, A. D. et al. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
22. Sun, W. et al. A map of the inorganic ternary metal nitrides. *Nat. Mater.* **18**, 732–739 (2019).
23. Li, H. et al. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE International Conference on Data Mining* 899–904 (IEEE, 2014).
24. Fusilier, D. H., Montes-y-Gómez, M., Rosso, P. & Cabrera, R. G. Detecting positive and negative deceptive opinions using PU-learning. *Inf. Process. Manag.* **51**, 433–443 (2015).
25. Zeng, X., Zhong, Y., Lin, W. & Zou, Q. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform.* **21**, 1425–1436 (2020).
26. Mordelet, F. & Vert, J.-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit. Lett.* **37**, 201–209 (2014).
27. Jang, J. et al. Structure-based synthesizability prediction of crystals using partially supervised learning. *J. Am. Chem. Soc.* **142**, 18836–18843 (2020).
28. 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).
29. Wang, J. & Saligrama, V. In *Local Supervised Learning through Space Partitioning, 2012* (eds. Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q.) (Curran Associates, Inc., 2012).
30. Oiwa, H. & Fujimaki, R. Partition-wise linear models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* Vol. 2, 3527–3535 (MIT Press, 2014).
31. Jung, E. H. et al. Efficient, stable and scalable perovskite solar cells using poly(3-hexylthiophene). *Nature* **567**, 511–515 (2019).
32. Bai, S. et al. Planar perovskite solar cells with long-term stability using ionic liquid additives. *Nature* **571**, 245–250 (2019).
33. Kojima, A., Teshima, K., Shirai, Y. & Miyasaka, T. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* **131**, 6050–6051 (2009).

34. Luo, J. et al. Efficient and stable emission of warm-white light from lead-free halide double perovskites. *Nature* **563**, 541–545 (2018).
35. Yuan, M. et al. Perovskite energy funnels for efficient light-emitting diodes. *Nat. Nanotechnol.* **11**, 872–877 (2016).
36. Lin, K. et al. Perovskite light-emitting diodes with external quantum efficiency exceeding 20 percent. *Nature* **562**, 245–248 (2018).
37. Zemen, J., Gercsi, Z. & Sandeman, K. G. Piezomagnetism as a counterpart of the magnetovolume effect in magnetically frustrated Mn-based antiperovskite nitrides. *Phys. Rev. B.* **96**, 024451 (2017).
38. Lukashev, P., Sabirianov, R. F. & Belashchenko, K. Theory of the piezomagnetic effect in Mn-based antiperovskites. *Phys. Rev. B.* **78**, 184414 (2008).
39. Oudah, M. et al. Superconductivity in the antiperovskite Dirac-metal oxide $Sr_3-x$SnO. *Nat. Commun.* **7**, 13617 (2016).
40. He, T. et al. Superconductivity in the non-oxide perovskite $MgCNi_3$. *Nature* **411**, 54–56 (2001).
41. Zhao, Y. & Daemen, L. L. Superionic conductivity in lithium-rich anti-perovskites. *J. Am. Chem. Soc.* **134**, 15042–15047 (2012).
42. Lü, X. et al. Antiperovskite $Li_3OCl$ superionic conductor films for solid-state Li-ion batteries. *Adv. Sci.* **3**, 1500359 (2016).
43. Kim, I., Do, J., Kim, H. & Jung, Y. Charge-transfer descriptor for the cycle performance of β-$Li_2MO_3$ cathodes: Role of oxygen dimers. *J. Mater. Chem. A* **8**, 2663–2671 (2020).
44. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
45. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
46. Liu, H. et al. Screening stable and metastable $ABO_3$ perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **177**, 109614 (2020).
47. Balachandran, P. V. et al. Predictions of new $ABO_3$ perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
48. Lu, S. et al. Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning. *Small Methods* **3**, 1900360 (2019).
49. Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 19 (2016).
50. Halder, A., Ghosh, A. & Dasgupta, T. S. Machine-learning-assisted prediction of magnetic double perovskites. *Phys. Rev. Mater.* **3**, 084418 (2019).
51. Xu, Q., Li, Z., Liu, M. & Yin, W.-J. Rationalizing perovskite data for machine learning and materials design. *J. Phys. Chem. Lett.* **9**, 6948–6954 (2018).
52. Shannon, R. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. Sect. A* **32**, 751–767 (1976).
53. Travis, W. et al. On the application of the tolerance factor to inorganic and hybrid halide perovskites: A revised system. *Chem. Sci.* **7**, 4548–4556 (2016).
54. Mochizuki, Y. et al. Theoretical exploration of mixed-anion antiperovskite semiconductors M3XN(M = Mg, Ca, Sr, Ba; X = P, As, Sb, Bi). *Phys. Rev. Mater.* **4**, 044601 (2020).
55. Sutton, C. et al. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 4428 (2020).
56. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
57. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
58. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
59. Chen, C. et al. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
60. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
61. Jena, A. K., Kulkarni, A. & Miyasaka, T. Halide perovskite photovoltaics: Background, status, and future prospects. *Chem. Rev.* **119**, 3036–3103 (2019).
62. Zhu, J. et al. Enhanced ionic conductivity with $Li7O_2Br_3$ phase in $Li_3OBr$ anti-perovskite solid electrolyte. *Appl. Phys. Lett.* **109**, 101904 (2016).
63. Asano, K., Koyama, K. & Takenaka, K. Magnetostriction in $Mn_3CuN$. *Appl. Phys. Lett.* **92**, 161909 (2008).
64. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
65. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
66. Perdew, J. P. et al. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).
67. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
68. Kresse, G. & Furthmüller, J. Software VASP, Vienna (1999). *Phys. Rev. B* **54**, 169 (1996).
69. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
70. Blöchl, P. E., Jepsen, O. & Andersen, O. K. Improved tetrahedron method for Brillouin-zone integrations. *Phys. Rev. B* **49**, 16223 (1994).
71. Stadelmaier, H. H. & Fraker, A. C. Stickstofflegierungen der T-Metalle Mangan, Eisen, Kobalt und Nickel mit Gallium, Germanium, Indium und Zinn. *Z. f.ür. Metallkd.* **53**, 48–51 (1962).
72. Holleck, H. The effect of carbon on the occurrence of $Cu_3Au$-type phases in actinide-and lanthanide-platinum metal systems. *J. Nucl. Mater.* **42**, 278–284 (1972).

## ACKNOWLEDGEMENTS

We acknowledge generous financial support from NRF Korea (NRF-2019M3D3A1A01069099). CPU and GPU resources from the Korea Institute of Science and Technology Information are greatly acknowledged.

## AUTHOR CONTRIBUTIONS

G.G., J.J., and J.N. contributed equally to this work. Y.J. conceived the idea. G.G. and J.N. collected the crystal data. G.G. and J.J. designed the machine learning framework. G.G. and J.J. searched the literature for the previous cases of virtual crystal synthesis. J.N. performed the DFT calculations. All authors discussed the results and assisted with the manuscript preparation.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00757-z.

**Correspondence** and requests for materials should be addressed to Yousung Jung.

**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.