Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00065c

Solid-state synthesizability predictions using positive-unlabeled learning from human-curated literature data†

Vincent Chung, *[a] Aron Walsh [a] and David J. Payne [abc]

The rate of materials discovery is limited by the experimental validation of promising candidate materials generated from high-throughput calculations. Although data-driven approaches, utilizing text-mined datasets, have shown some success in aiding synthesis planning and synthesizability prediction, they are limited by the quality of the underlying datasets. In this study, synthesis information of 4103 ternary oxides was extracted from the literature, including whether the oxide has been synthesized *via* solid-state reaction and the associated reaction conditions. This dataset provides an opportunity to supplement existing solid-state reaction models *via* reliable data and information from articles whose content and formats are challenging to extract automatically. A simple screening using this dataset identified 156 outliers from a subset of a text-mined dataset that contains 4800 entries, of which only 15% of the outliers were extracted correctly. Finally, this dataset was used to train a positive-unlabeled learning model to predict the solid-state synthesizability of new ternary oxides, where we predict 134 out of 4312 hypothetical compositions are likely to be synthesizable.

Received 16th February 2025
Accepted 16th July 2025

DOI: 10.1039/d5dd00065c

rsc.li/digitaldiscovery

1 Introduction

In recent years, with the improvement of software and the enhanced availability of materials data from projects such as the Materials Project[1] and the Inorganic Crystal Structure Database (ICSD),[2] high-throughput screening of hypothetical materials has become possible *via* theoretical calculations and data-driven methods. However, the bottleneck of materials discovery is mainly found in the experimental validation, which often relies on trial-and-error approaches. Therefore, researchers have started to look for new methods to predict the synthesizability of hypothetical compounds.[3] A popular metric to serve as a proxy to material synthesizability is the thermodynamic stability metric, energy above the convex hull (E_{hull}), which is defined as the difference between the formation enthalpy of the material and the sum of the formation enthalpies of the combination of decomposition products that maximize the sum.[4] It has been used extensively to filter different hypothetical materials,[5–9] however, E_{hull} is not a sufficient condition for synthesizability since a non-negligible number of hypothetical materials with low E_{hull} have not been synthesized.[10] Since E_{hull} is often calculated from internal energies at 0 K and 0 Pa, the actual thermodynamic stability of the material will vary depending on the synthesis conditions. In addition, E_{hull} does not account for kinetic factors, where kinetic barriers could prevent an otherwise energetically favorable reaction or phase change from occurring. A well-known example is martensite, which is commonly synthesized by quenching austenite. E_{hull} also does not take into account the entropic contribution to materials stability.[11] Although theoretical approaches have enabled a detailed understanding of the synthesis routes, for example the crystallization pathway of MnO_2 during hydrothermal synthesis,[12] they are currently computationally unfeasible to apply them in high-throughput studies for thousands of materials.

A promising and scalable method is to use data-driven approaches to learn from synthesis records. Raccuglia *et al.* used hydrothermal reaction data from their laboratory notebooks to train a model to predict the reaction outcome of templated vanadium selenite.[13] Bartel *et al.* applied the independence screening and sparsifying operator (SISSO) method to analyze synthesized perovskite oxides and halides to create a new tolerance factor, with overall improved performance compared to the traditional perovskite Goldschmidt tolerance factor.[14,15]

The first major obstacle for data-driven approaches to predicting materials' synthesizability is the low quantity and quality of relevant data. Synthesis information is not easily accessible on a large scale because it is commonly stored in text

[a]Department of Materials, Imperial College London, South Kensington, London SW7 2AZ, UK. E-mail: vincent.chung15@imperial.ac.uk

[b]Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA, UK

[c]NEOM Education, Research, and Innovation Foundation, Al Khuraybah, Tabuk 49643-9136, Saudi Arabia

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5dd00065c

format in the literature or private lab books.[13] To overcome this challenge, natural language processing (NLP) techniques have been used to build material synthesis datasets. Kim *et al.* developed an autonomous framework to extract the synthesis parameters from over 640 000 journal articles on 30 oxide systems.[16] This was later expanded and improved, and the resulting text-mined dataset was used to predict hydrothermal synthesis conditions[17] and to plan the solid-state synthesis of metal oxides.[18] Later on, Kononova *et al.* developed a text mining pipeline for solid-state reactions and sol–gel synthesis data from the literature,[19] which was used to train a reaction graph model for predicting the major product[20] and synthesis conditions[21,22] of solid-state reactions. More recently, models trained with text-mined datasets were used to generate synthesis recipes for an autonomous laboratory that accelerated the discovery of novel materials.[23]

As data availability increases, the bottleneck of data-driven approaches shifts from quantity to quality of text-mined datasets. As demonstrated in the chemistry community, the difference in performance between a well-filtered and noisy dataset should not be ignored.[24–26] The overall accuracy of the Kononova *et al.* dataset (where all of the extracted synthesis conditions and actions of the entry are correct) is only 51%,[19] which was cited by Malik *et al.* as a reason to use coarse descriptions of synthesis action (*e.g.* mix/heat/cool) instead of detailed descriptions (*e.g.* heating temperature/time) in their study.[20] While it is widely acknowledged that the quality of text-mined datasets is lower than their manual counterparts,[27] no quantitative analysis between the two has been performed in the materials domain, which could have served as a milestone or a goal for the material text-mined dataset.

Another issue with current material synthesis data, as highlighted by Raccuglia *et al.* and Jensen *et al.*, is that it is rare for papers to include failed material synthesis attempts,[13,28] which is challenging to resolve without a change in the scientific community. One approach to overcome the lack of failed reaction data is positive-unlabeled (PU) learning, a type of semi-supervised learning when only positive and unlabeled data are available.[29] Frey *et al.* adopted a transductive bagging PU learning approach developed by Mordelet *et al.* to predict the synthesizability of 2D MXene and their precursors.[30,31] Jang *et al.* later used a similar approach to predict the synthesizability of hypothetical compounds in the Materials Project.[32] Recently, Gu *et al.* used inductive PU learning and domain-specific transfer learning to predict the synthesizability of general perovskites, which showed better performance than Jang *et al.*'s and tolerance factor-based approaches.[33] However, all three studies can only evaluate the positive data, so it is difficult to estimate the number of false positives (compounds that cannot be synthesized but are classified as synthesizable).

In this paper, a dataset that contains information on whether ternary oxides in the Materials Project database with ICSD IDs have been synthesized *via* solid-state reaction was human curated (*i.e.* manually). This included articles whose formats are difficult for automatic extraction. Potential applications of this human curated dataset are illustrated in the following sections: (1) analysis of E_{hull} with solid-state synthesizability, defined as whether the material can be synthesized *via* solid-state reaction, as opposed to general synthesizability; (2) outlier detection of a text-mined dataset on solid-state reaction; (3) prediction of solid-state synthesizability using a PU learning framework.

2 Experimental

2.1 Data collection

The manual data collection was performed by the first author, who had prior experience with solid-state synthesis. Firstly, 21 698 ternary oxide entries were downloaded from the Materials Project[1] (version 2020-09-08) *via* pymatgen.[34] Next, by using the ICSD IDs as an initial proxy of synthesized materials, 6811 entries with at least one ICSD ID were identified. Afterwards, entries with non-metal elements and silicon were removed, resulting in 4103 ternary oxide entries (with 3276 unique compositions from 1233 chemical systems) for manual data extraction *via* ICSD, Web of Science, and Google Scholar. Explained briefly, the search process was as follows: (1) examination of the papers corresponding to the ICSD IDs; (2) examination of the first 50 search results sorted from oldest to newest in Web of Science with the chemical formula as input; (3) examination of the top 20 relevant search results in Google Scholar with the chemical formula as input. Additional information on the manual data collection process can be found in ESI S1.†

Each ternary oxide has been checked in the literature for whether it has been synthesized *via* a solid-state reaction. If there is at least one record that the compound has been synthesized *via* solid-state reaction, the highest heating temperature, pressure, atmosphere, mixing/grinding condition, number of heating steps, cooling process, precursors, and whether the synthesized product is single-crystalline, were collected when available. Otherwise, the material would be labeled as non-solid-state synthesized (the material has been synthesized but not *via* solid-state reactions). For entries in which there was insufficient evidence that the ternary oxides have been synthesized *via* solid-state reactions, they were labeled as undetermined. The reasons for these undetermined entries are provided in the comment section of the dataset. In total, the human curated dataset contains 3017 solid-state synthesized entries, 595 non-solid-state synthesized entries, and 491 undetermined entries. The data description of the human curated dataset can be found in ESI S2.†

For comparison, Kononova *et al.*'s text-mined solid-state reaction dataset (ver. 2020-07-13) was downloaded through their repository.[19] It contains 31 782 solid-state reaction entries from the literature after the year 2000. The definition of whether a synthesis is considered a solid-state reaction in this study differs from those in previous studies. Huo *et al.* defined a solid-state reaction as follows: (1) the input materials are subjected to a process of grinding/milling; (2) the powders are mixed and heated.[35] During data curation, we observed that a non-negligible number of papers did not explicitly state the grinding/milling steps, so the first criteria was dropped. We also added two additional criteria for a synthesis to be considered a solid-state reaction: (3) the reaction does not involve flux or

cooling from melt (except for high-pressure solid-state synthesis where oxidizers were used with a secondary function as flux or mineralizer for higher crystallinity when explicitly stated); (4) the heating temperature must not be above the melting point of all the starting materials. Details on the processing of the text-mined dataset can be found in ESI S3.1.†

For the analysis of the dataset, binary oxide melting points were taken from the CRC Handbook of Chemistry and Physics online[36] and other papers. Some of the melting points are the decomposition temperature or the transition temperature.

2.2 Data validation

The method of evaluation of the human curated dataset depends on the labeling of the entries. For solid-state synthesized entries, 100 randomly chosen entries of different compositions were reexamined by checking against the entries' references. For non-solid-state synthesized entries, 55 randomly chosen compositions (10% of the non-solid-state synthesized compositions) were reexamined following the same procedure as manual data curation.

The metrics used for the validation were recall and precision. The recall and precision used in the paper were based on the definitions used by Raghavan and Jung for information retrieval.[37] The recall is the ratio of the number of correctly extracted values to the number of relevant values. The precision is the ratio of correctly extracted values to the number of extracted values. For consistency, data validation was performed by the same (first) author who manually extracted the dataset. The formula and an example of the recall and precision calculation are shown in ESI S4.†

2.3 Data preprocessing for PU learning

To apply PU learning to predict the solid-state synthesizability of ternary oxide compositions, three types of data are required: (1) solid-state synthesized compositions, which are considered as solid-state synthesizable compositions; (2) non-solid-state synthesized compositions (compositions that have been synthesized but not *via* solid-state reaction), which are assumed to be solid-state unsynthesizable compositions (compositions that cannot be synthesized *via* solid-state reactions but can be synthesized *via* other methods); (3) hypothetical compositions, which contain solid-synthesizable, solid-state unsynthesizable, and unsynthesizable compositions.

The solid-state synthesized and non-solid-state synthesized compositions were gathered from the human curated dataset, while the hypothetical compositions were collected from the Materials Project ternary oxide entries without ICSD IDs or with ICSD IDs that reference a computational (*i.e.* non-experimental) paper. All compositions were then featurized using matminer[38] and basic mathematics operations based on their binary oxide melting points (ESI S5.1†). Compositions that contain the element Pa or have difficulties in oxidation state assignment were removed, resulting in 7033 compositions in the dataset for PU learning. This dataset for PU learning contains 2213 solid-state synthesized, 508 non-solid-state synthesized, and 4312 hypothetical compositions.

2.4 PU-learn data labeling and model training

The PU learning method used in this study is a modification of the decision tree adoption code from Frey *et al.*,[31] which originated from the transductive bagging support vector machine scheme proposed by Mordelet *et al.*[30] Each iteration of positive-unlabeled learning is as follows: let **P**, **N**, **U** be the positive, negative, and unlabeled data sets, with **K** and **H** being the number of positive and negative data points in **P** and **N**, respectively. At the beginning of each iteration, **K**–**H** number of data points in **U** are randomly labeled as negative to ensure balanced classes. This subset of **U**, denoted as $\mathbf{U_n}$, is then used to train a decision tree classifier along with **P** and **N**. The trained classifier will predict the solid-state synthesizability of the remaining unlabeled data points in **U** ($\mathbf{U}\backslash\mathbf{U_n}$) with a score between 0 and 1. One hundred iterations with different $\mathbf{U_n}$ are repeated so that the solid-state synthesizability score of each data point is the average of its scores in the iterations where it is not in $\mathbf{U_n}$ (out-of-sample score). The evaluation of the model is carried out with 10-fold cross-validation and averaging the out-of-sample scores for each data point. To minimize the effect of data splitting on model evaluation, 10-fold cross-validation was repeated 10 times with different data splits.

In total, three PU learning models were trained based on different labeling schemes but on the same dataset, as shown in Fig. 1. The task of model 1 and model 2 is the prediction of solid-state synthesizability, while the task of model 3 is the prediction of general synthesizability. The data labeling schemes for the three models are as follows:

- For model 1, the solid-state synthesized compositions are positively labeled, while the hypothetical compositions and non-solid-state synthesized compositions are unlabeled.
- For model 2, the solid-state synthesized compositions are positively labeled, the non-solid-state synthesized compositions are negatively labeled, and the hypothetical compositions are unlabeled.
- For model 3, the solid-state synthesized and non-solid-state synthesized compositions are positively labeled, while the hypothetical compositions are unlabeled.

The data labeling of models 1 and 2 differ in the labeling of non-solid-state synthesized compositions, where they are unlabeled in model 1 and negatively labeled in model 2. Two schemes were tested and compared because the non-solid-state synthesized compositions are relatively noisier and not a random representation of solid-state unsynthesizable compositions (*e.g.* unsynthesizable compositions are also solid-state unsynthesizable compositions, but not non-solid-state synthesized compositions).

In addition to the three PU learn models, a supervised learning Lightgbm[39] classification model was trained without using the hypothetical compositions to show why PU learning is required for synthesizability prediction. For this supervised learning model, the solid-state synthesized compositions are positively labeled and the non-solid-state synthesized compositions are negatively labeled. The data was separated into train-test sets with an 8 : 2 ratio in a stratified manner. 10-Fold cross-validation was then performed on the training set. In the end,

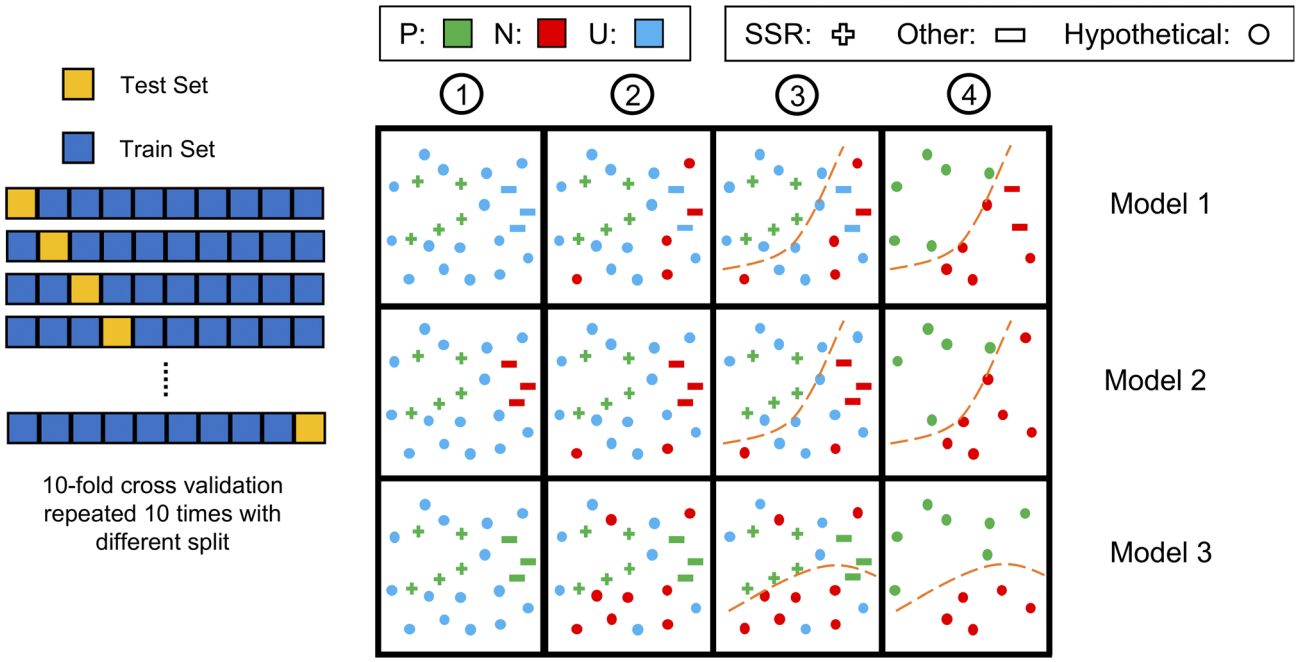

Fig. 1 Illustration of the transudative bagging PU learning process of model 1, 2, and 3 to predict the solid-state synthesizability (model 1 and 2) or the general synthesizability (model 3) of hypothetical compositions. For each iteration, the 4 steps are as follows: (1) initialization of labeling based on the model (2) random labeling of unlabeled data points until the number of positively and negatively labeled data points are equal (3) classification based on positively and negatively labeled data points (4) prediction of the remaining unlabeled data points based on the classifier trained in step 3. The plus, rectangle, and circle dots in the illustration represent the compositions synthesized *via* solid-state reaction (SSR), compositions synthesized by other non-solid-state reaction methods (other), and hypothetical compositions (hypothetical), respectively.

model evaluation was performed on the test set. The Lightgbm model was trained with an AMD Radeon GPU and specified the use of the 64-bit float point setting to prevent reproducibility issues experienced when using a 32-bit float point (the default) for number summations.[39] Details of feature selection and hyperparameter tuning of all models are in ESI S5.2 and S5.3.†

2.5 Model evaluation and metric

The area under the receiver operating characteristic curve (ROC AUC) was chosen as the metric for model tuning and evaluation. The decision threshold of each model was chosen such that it maximizes the geometric mean (*G*-mean) between the true positive rate (TPR) and the true negative rate (TNR, equivalent to 1 – false positive rate (FPR)). This was chosen to emphasize the balance in classifying both solid-state synthesizable and solid-state unsynthesizable compositions correctly.

For model 3, due to the absence of negative data (unsynthesizable compositions), the bias ROC AUC was computed instead, which assumes all unlabeled data (hypothetical compositions) as unsynthesizable compositions and labeled as negative. For a clean positive data set (no mislabeled positive entries), the relationship between the true AUC and the biased AUC (AUC^{PU}) is described by the following equation:[40]

$$AUC = \frac{AUC^{PU} - \frac{\alpha}{2}}{1 - \alpha}$$

where α is the ratio of positive samples in the unlabeled data set and the AUC^{PU} is the biased AUC score. Although the true value of α is unknown, AUC and AUC^{PU} are monotonic when α is fixed, so it can be a proxy for AUC for model evaluation.

3 Results and discussions

3.1 Evaluation of the human curated dataset

The precision of solid-state synthesized labels is 0.99 and 0.86 for solid-state synthesized and non-solid-state synthesized entries, respectively. The large difference is due to the variation in the requirement: an examination of only one paper is required to verify whether a solid-state synthesized label is correct, but the whole manual collection process was repeated to verify that a material has not been synthesized *via* solid-state reactions in the literature.

The precision and recall of the solid-state reaction conditions extracted in the solid-state synthesized entries are shown in Table 1. Overall, the precision of the extracted conditions is high (0.96–1), while the recall is slightly lower (0.89–1). Lower extraction performance was observed for columns that contain more information, namely the cooling and mixing/grinding conditions.

3.2 Comparison with text-mined dataset

3.2.1 Comparison of extraction performance. Kononova *et al.*'s text-mined dataset contains 28 604 solid-state synthesis recipes extracted from 21 794 papers,[19] whereas the human

Table 1 Error estimation of manual data extraction from the literature

Metric	Solid-state synthesized	Precursor	Heating temperature	Heating pressure & atmosphere	Cooling	Mixing/grinding
Precision	0.99	0.96	0.98	1	1	0.96
Recall	N/a	0.99	1	0.98	0.9	0.89

curated dataset contains 2295 solid-state synthesis entries extracted from 1874 papers, of which 154 papers are present in both datasets. The text-mined dataset includes sequences of synthesis steps (mixing and heating conditions), whereas the human curated dataset only includes certain synthesis conditions (*e.g.* highest heating temperature, cooling conditions). Due to the difference in the datasets, 100 randomly selected entries from the text-mined dataset were examined and compared with the human curated dataset in three categories: whether the target was synthesized *via* solid-state reaction, whether the correct precursors were collected, and whether the highest heating temperature is correct (Table 2). Out of the 100 randomly examined entries, the following were observed for the three categories:

• Target material synthesized *via* non-solid-state methods:

– Three of them were synthesized *via* methods that do not fit Huo *et al.*'s definition of a solid-state reaction.[35] These alternative synthesis methods are sol–gel precursor synthesis, solid–gas reaction, and mechanical activation of precursors without heating.

– Two of them described the solid-state synthesis attempts of $Nd_6Mo_{10}O_{39}$ and $LiBiO_3$, but according to their respective result and discussion sections, their synthesis attempts were unsuccessful.[41,42]

– Seven of them had erroneous target strings. In some of these cases, the target composition described the nominal composition of the mixture instead of the actual composition of the target.

• Incorrect precursor:

– Three out of seven erroneous extracted precursors were due to the extraction of precursors from another target or synthesis route in the same paragraph.

• Incorrect highest heating temperature:

– Eleven of them did not contain the heating step with the highest heating temperature.

– Ten of them did not extract the correct highest heating temperature from the heating step.

The recall of the precursor category for the text-mined dataset is not applicable because Kononova *et al.* filtered out entries where balanced chemical reactions could not be formed using the target and precursors of the entries.[19] Therefore, all of the text-mined entries have precursor information. Out of the three categories, the highest heating temperature performed the worst, with a precision and recall of 0.85 and 0.87, respectively, which are notably worse than the human curated dataset's precision and recall.

The overall lower extraction performance suggests that a model trained using the text-mined dataset should perform worse than a model trained using the human curated dataset. However, the severity of the impact would require further investigation. For example, solid-state reactions can occur over a relatively wide temperature range and the range can be several times larger than the mean absolute error (MAE) of the highest heating temperature in the text-mined dataset, which was calculated to be 48 °C. On the other hand, models for synthesis condition prediction in the literature used features generated from the target and precursor information.[21] Therefore, the extraction errors of the precursor and target should be considered. Other factors, such as the number of chemical systems, are also important and can affect the generalization of models. For the 100 entries in the text-mined dataset examined, only 63% of the entries have the correct solid-state synthesized target, precursors, and highest heating temperature.

3.2.2 Comparison of method. Aside from the improvement of data quality, an advantage of human curation is that it can be performed on older papers that are not readily available in a wide range of digital formats but often only in PDF, which is more challenging for automatic extraction. This was why Kononova *et al.* extracted from papers published in HTML/XML only.[19] This also applies to supplementary information that is often unstructured and therefore difficult for automatic extraction.[28]

Another limitation of the current automatic data extraction is that many of them focus only on the paragraphs describing the synthesis, which might ignore relevant information located in other sections of the paper. The focus on extraction from a single paragraph in the methods section also misses opportunities to curate information on the target material such as crystal structure and physical properties, which would provide useful information for future investigations on wider quantitative structure–property relationships. However, training a model to collect information from non-method sections of the paper increases training complexity and time, so this trade-off has to be considered. Human curation also offers increased flexibility, where additional but crucial information or details outside of the initial extraction plan can be collected.

The largest disadvantage of human data extraction is the time required to curate the data. Data collection of the 4103 ternary oxide entries took around a year for a single person, with synthesis condition extractions taking up to 10 minutes per

Table 2 Error estimation of the text-mined dataset

Metric	Solid-state synthesized	Precursor	Heating temperature
Precision	0.88	0.93	0.85
Recall	N/a	N/a	0.87

paper. This is significantly longer than the automatic pipelines for text mining synthesis conditions as once the models are trained, the time it takes per paper is on the scale of seconds. A possible circumvent is through a combination of crowd-sourcing and semi-automated text mining methods,[43] but there are some difficulties to overcome, mainly the software for collaborated annotation and the lack of community agreed standards.[27] In addition, once the automatic pipeline is trained, the information extracted from the same paper will be consistent and the decision-making process can be reviewed and analyzed. The same cannot be done for manual data extraction readily, even if strict criteria are applied.

3.3 Analysis of the human curated dataset

To date, this dataset is the largest human curated dataset on the solid-state synthesis information of known ternary oxides. Although not all synthesis steps and conditions were recorded, it allows analysis of the relationship between synthesis, structure, and properties of materials.

3.3.1 E_{hull} and synthesizability. Fig. 2 shows the synthesized ratio and solid-state synthesized ratio of ternary oxides in the Materials Project at different E_{hull} intervals. The synthesized ratio is the number of synthesized ternary oxides over the number of all ternary oxides (synthesized and hypothetical), while the solid-state synthesized ratio is the number of solid-state synthesized ternary oxides over the number of synthesized ternary oxides. In other words, the solid-state synthesized ratio represents the proportion of all synthesized materials that can be synthesized *via* solid-state reactions.

Fig. 2a agrees with the previous analysis of computational material databases where the frequency of synthesized materials decreases at higher E_{hull},[10,44] making it a simple criterion to account for the likelihood of finding synthesizable materials. In addition, we found that around 6% of the examined ternary oxide entries are hydrates in the ICSD but not in the Materials Project, where only hydrogen atoms are missing. Most of them tend to have higher energy (81.3% of them have an E_{hull} above 100 meV per atom) so setting a heuristic upper limit to E_{hull} during screening may be effective at filtering them out.

However, the ratio of solid-state synthesized materials and synthesized materials does not follow the same trend, where the ratio is 0.69–0.87 across all E_{hull} values (narrower E_{hull} intervals show the same trend, as shown in ESI S6†). This suggests that it may not be possible to set a simple energy threshold to determine the solid-state synthesizability. There are a few reasons:

- Solid-state reactions have a wide range of reaction conditions that could change throughout the synthesis process. This means that calculation of the free energy of reactions at one particular set of reaction conditions and reactants cannot reflect the stability of phases throughout the reaction, for example, because of the formation of intermediate phases.
- At temperature typically used in solid-state reaction, entropic contribution to the free energy of competing phases and reactions cannot be ignored.[11] In addition, configurational entropy also plays a major role in the stability of multi-component materials.[45,46]
- The kinetic barriers of reactions have an effect on the formation of intermediate phase during solid-state reaction, which could affect the final synthesis target. Todd *et al.* show that for the metathesis reaction $LiMnO_2$ + YOCl → LiCl + $YMnO_3$, using different polymorphs of $LiMnO_2$ as the reactant will alter the rate of the initial reaction, resulting in the formation of different polymorphs of $YMnO_3$.[47] Non-equilibrium cooling from high temperature can also stabilize metastable phases and prevent their decomposition. For example, the crystal structure of solid-state synthesized

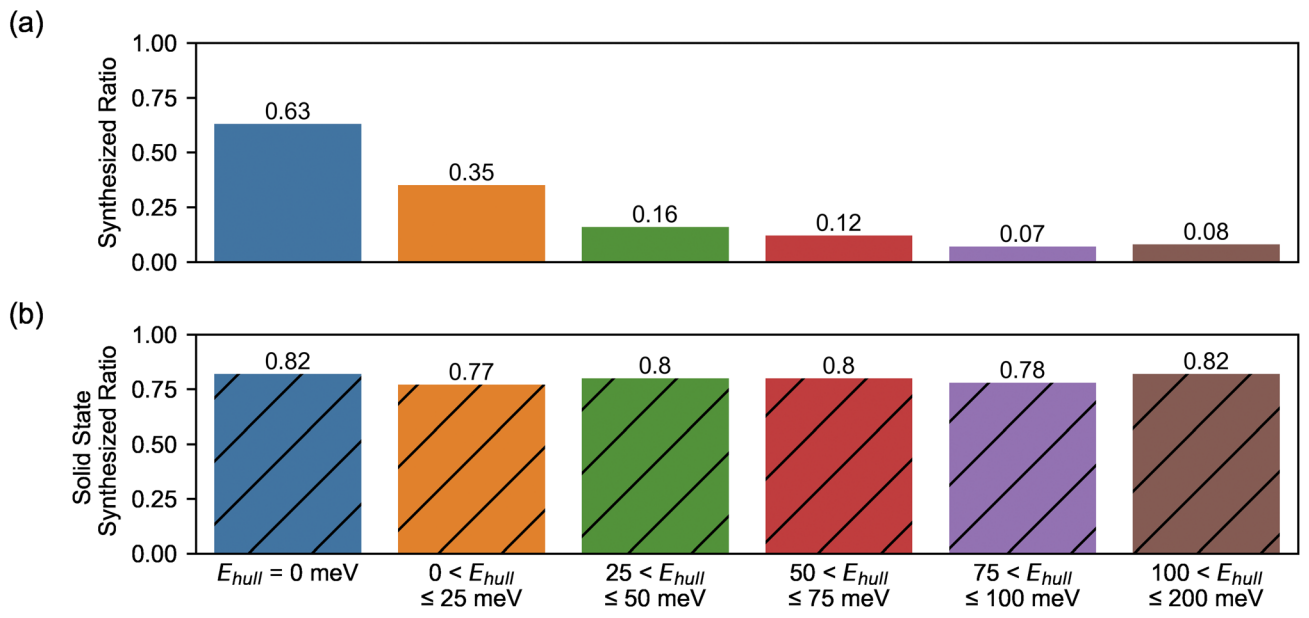

Fig. 2 The (a) synthesized ratio and (b) solid-state synthesized ratio of Materials Project ternary metal oxides at different E_{hull} intervals. Broader E_{hull} intervals were used for higher E_{hull} as there are fewer entries.

$Bi_{0.4}Pb_{0.6}Mn_{0.4}Ti_{0.6}O_3$ depends on whether the reaction was quenched or slowly cooled.[48]

3.3.2 Solid-state synthesis condition. The solid-state synthesized ratio differs between the elements in the ternary oxides (Fig. 3), from 33% for Se ternary oxides to 94% for Y ternary oxides, with an average of 80% for all ternary oxides. One reason is the relatively low melting point of the constituent oxide melting point, where the precursor is too volatile at high temperatures during a solid-state reaction. This is supported by the fact that the solid-state reaction temperatures used to synthesize the ternary oxides of Se, As, Hg, Ag, and Te are among the lowest (ESI S7†).

Another observation is that ~96% of the entries have at least one binary oxide as one of the precursors, and ~58% of them used binary oxides exclusively. Similar observations were observed by He *et al.* for the text-mined dataset, where simple oxides are often the most commonly used precursor, attributed to their stability under ambient conditions.[49] The dominant usage of a few types of precursor implies that the trends in synthesis condition obtained from an analysis of literature data are biased. This should be taken into consideration by researchers when using this information for the prediction of synthesis condition.

Compared to heating conditions, which are mostly reported numerically, the descriptions of cooling conditions can be either numerical (either the cooling rate or time), descriptive (*e.g.* furnace cooling, slow cooling), or both, with descriptive conditions being more common. The frequency of the three most common descriptions of cooling conditions (quench, slow cooling, and furnace cooling) and the cooling condition information (cooling rate, cooling time, and cooling medium) are shown in Fig. 4.

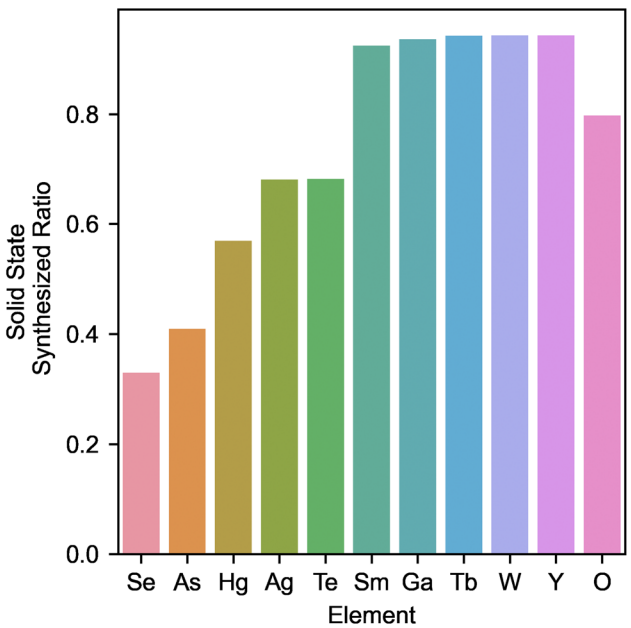

Fig. 3 The solid-state synthesized ratio of the 10 selected elements with the highest or lowest ratio that have more than 50 synthesized ternary oxides in the human curated dataset. Oxygen represents the average of all ternary ratios in the dataset.

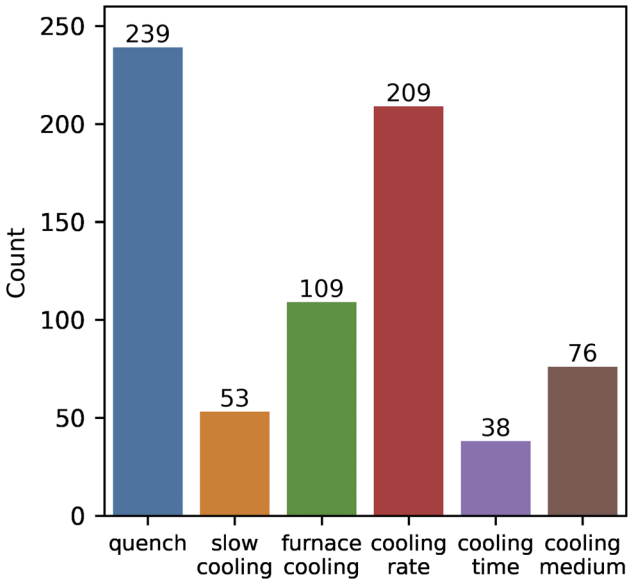

Fig. 4 Frequency of cooling condition descriptions and information out of the 607 entries that have cooling conditions in the human curated dataset.

The choice of cooling conditions, similar to the choice of precursors, is also influenced by researchers' bias. For example, out of the 607 examined papers with cooling conditions in the human curated dataset, 544 (90%) have one cooling step while 63 (10%) have two or more cooling steps. In particular, out of 47 Mo ternary oxide entries, 29 of them used two or more cooling steps for the solid-state synthesis, which is a much higher ratio than other ternary oxides'. An examination of the entries reveals that the reason is likely due to the researcher's bias, since 26 out of 29 entries came from 18 reference articles with at least one but often two common authors.

On the other hand, descriptions of mixing and grinding conditions are mostly descriptive. As shown in Fig. 5, manual mixing and grinding are more popular than mechanical alternatives (*e.g.* ball milling and vibration milling). The top three liquid mediums used for wet milling are acetone, alcohol (mainly ethanol), and water. The high usage rate of these liquids is probably due to the fact that these chemicals are widely available in all laboratories, and common precursors used in solid-state synthesis are insoluble in these liquid mediums. Although the choice of liquid mediums may not affect the chemical composition of the synthesis target, different liquid mediums would influence the particle size and geometry of the precursors, which in turn would affect the properties of the synthesis product.[50,51]

3.3.3 Reporting trend of solid-state synthesis conditions. The completeness of the descriptions of the synthesis conditions varies between the papers and the type of information. The precursor and heating temperature are almost always explicitly stated in the paper or referred to in another referenced paper, whereas information about the atmosphere, pressure,

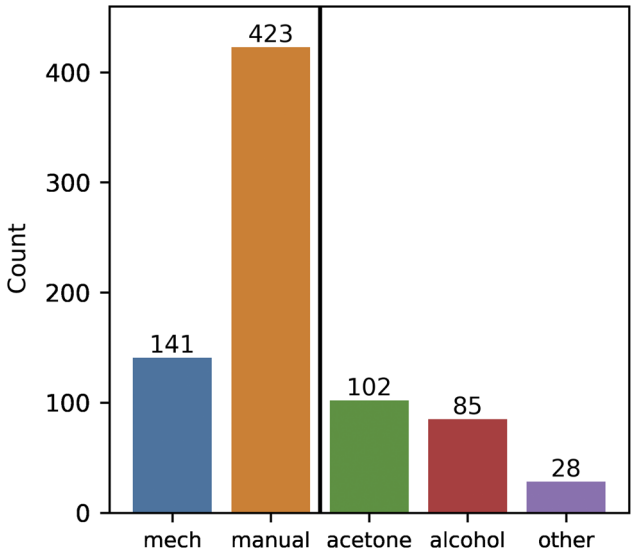

Fig. 5 Frequency of mix/grind condition descriptions and information out of the 561 entries that have mix/grind condition in the human curated dataset. The left of the black line is whether the precursors are mixed/ground manually or mechanically, while the right of the black line is the type of liquid medium used for wet milling.

grinding, and cooling conditions is often omitted. For cooling and mixing/grinding conditions, only 607 (27%) and 561 (25%) entries have any recorded information, respectively, which is only around a quarter of the number of entries with heating temperature and precursor information. Information such as the cooling medium has an even lower rate of being included in papers, as shown in Fig. 4. This is possibly because cooling and grinding/mixing conditions are considered less influential on most solid-state reaction outcomes in comparison to the precursor and heating conditions. Another possible reason is that some synthesis details are common knowledge or trivial, such that they are omitted.

Despite the lower reporting frequency of cooling and mixing conditions, they are sometimes important to solid-state synthesis, for example:

- The solid-state reaction between Er_2O_3 and Na_2CO_3 would yield $NaErO_2$ with either α-$NaFeO_2$ type or β-$LiFeO_2$ type structure depending on the cooling condition.[52]
- Wet milling of Bi_2O_3 precursors with acetone under ambient atmosphere would introduce carbon contaminates, which could lead to the formation of $Bi_2O_2CO_3$ as impurities in solid-state reactions.[53]
- Chen *et al.* compared solid-state synthesized $LiNi_{0.5}Mn_{1.5}O_4$ samples prepared with ball-milled and manually ground precursors and noticed a difference in electrochemical performance.[54]

Unfortunately, there is usually no clear indication of the importance or necessity of these synthesis conditions on the synthesis outcome.

Interestingly, the published date of the article affected the degree of omission of data. Out of 1149 entries that referenced pre-2000 sources (1929–1999), only 223 (19%) and 127 (11%) have any cooling and mixing conditions, respectively. This increased to 384 (26%) and 434 (30%) for cooling and mixing conditions, respectively, out of 1469 entries referenced from post-2000 sources. There are also minor differences in the reported synthesis conditions. For example, there are more papers that reported using furnace cooling or have defined cooling rates in post-2000s papers.

In summary, the reporting habit and frequency of the synthesis conditions from the literature should be taken into account when preparing an annotated synthesis text to train a text-mining algorithm to extract synthesis conditions. As discussed above, certain conditions like cooling and mixing conditions are reported less frequently and sometimes in greater varieties (*e.g.* cooling conditions could be a cooling rate, cooling time, or a description), which means a greater amount of annotated synthesis text is required for the model to learn to extract these conditions compared to more common conditions like the precursors and heating temperature.

3.4 Outlier detection

The simplest strategy to reduce the errors of text-mined datasets is to use a filter to select only high-quality data. Although this approach may reduce the amount of data, it can be readily applied to any dataset and is algorithm-independent. The use of balanced chemical reactions to filter good precursor and targets from Kononova *et al.*'s text-mined dataset is one example. The Synthesis Materials Recognizer (SMR) model has a precision and recall of 0.889 and 0.912 for precursors.[49] By filtering entries with balanced chemical reactions, both the precision and recall increased to 0.99 for the text-mined dataset.[19]

As a demonstration, the general trend of ternary oxide calcination temperature observed from the human curated dataset was applied to find outliers in a subset of the text-mined dataset. The text-mined subset consisted of 4800 entries where each entry contains one ternary metal oxide as the target material and the constituent elements of the target material are all present in the human curated dataset. The ratios of the highest heating temperature and the minimum/maximum/mean of the binary oxide melting point were chosen as filters. This choice was made because the melting points of the precursors are a common consideration of the heating temperature used for solid-state synthesis, *e.g.* Tammann-rule.[55] Binary oxide melting points were used in place of precursor melting points as the latter are not easily available, and binary oxides are used in most solid-state synthesis for ternary oxides.

The cumulative distribution of the ratio of the highest heating temperature and the maximum binary oxide melting point for the human curated dataset and text-mined subset is shown in Fig. 6, which was chosen due to the highest Pearson's correlation out of the three ratios (ESI S8†). 90% of the text-mined subset has a ratio between 0.44 and 0.80, which is narrower than the range of the human curated dataset (0.40 and 0.87). The wider distribution is due to a larger number of chemical systems in the human curated dataset (1072 *vs.* 665) despite having fewer entries (2234 *vs.* 4800). The inclusion of more chemical systems could improve the generalizability of

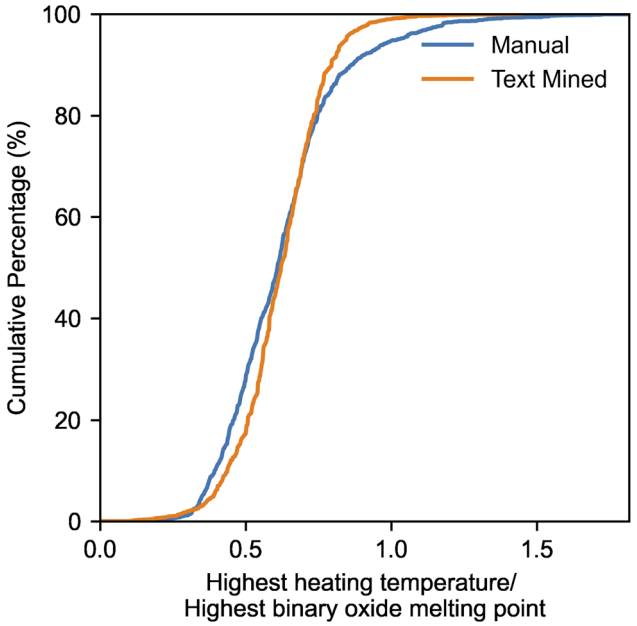

Fig. 6 The cumulative percentage of the ratio between the highest heating temperature and highest binary oxide melting point of the ternary oxides in the human curated dataset and the text-mined subset.

machine learning models, as shown by D. Jha *et al.* where the MAE of the predicted formation enthalpy of compounds increased when the compounds that share the same chemical systems were removed from the training data.[56]

To identify the outliers, the minimum/maximum and 1st/99th percentile pairs of the 3 ratios calculated from the human curated dataset were chosen as thresholds to filter the text-mined subset. The reference papers of the entries considered to be outliers were examined to verify whether the synthesis target, method, and highest heating temperature were correctly extracted. Discrepancies in the definition of solid-state synthesis were accounted for during examination by using Huo *et al.*'s definition.[35]

Out of the 52 entries identified as outliers using the minimum/maximum ratios as filters, only 5 (∼10%) were correctly and completely extracted. Whereas, when using 1st and 99th percentile as filters, the correct ratio increased to ∼15% (23/156). Most of these outliers were detected because their ratios are below the lower thresholds, whereas only 6 entries out of 156 have ratios above the upper thresholds.

Out of the 133 erroneous entries, around 23 (49%) have erroneous highest heating temperatures, 45 (34%) have the wrong synthesis method, and 23 (18%) have the wrong target. Most of these errors are described in the earlier sections on the evaluation of the text-mined dataset. In particular, it was observed that sentences that described heating operations but used a general verb that can apply to all synthesis methods (*e.g.*, "synthesis" and "prepare") were not extracted properly because the action was defined by the researchers as "starting operations" as opposed to "heating operations". This illustrates how a possible error in the data annotation could affect automatic text extraction. An example of the above can be found in ESI S3.2.†

The higher quality of the human curated allows the derivation of simple criteria for filtering text-mined datasets, which can reduce error propagation when the dataset is used for model training and also help researchers identify the limitations and flaws in the automatic text extraction process.

3.5 PU-learn

3.5.1 Evaluation. Fig. 7 shows the solid-state synthesizability scores of the synthesized ternary oxides in the Materials Project using model 1 and 2. The decision thresholds and performance of the models are shown in Table 3, where model 1 outperformed model 2 with a *G*-mean score of 0.863 with a much lower FPR. As the only difference between model 1 and model 2 is the initial labeling of non-solid-state synthesized compositions during model training (as they are unlabeled and negatively labeled for model 1 and model 2, respectively), the result suggests the initial negative labeling of non-solid-state synthesized compositions has led to a worse model performance. This can be explained by two reasons:

(1) The non-solid-state synthesized compositions cannot fully represent the solid-state unsynthesizable compositions, because they do not include unsynthesizable compositions, which are present only among the hypothetical compositions.

(2) As shown in previous sections, the probability of non-solid-state synthesized compositions mislabeled during manual data collection is relatively high (precision of 0.86), which means that around 14% of them should be positively labeled instead.

Therefore, in all iterations during training, model 2 classifiers learned from a higher proportion of non-solid-state synthesized compositions that are negatively labeled (step 3 in Fig. 1), which reduces the effectiveness of the model in learning from the unlabeled data and in distinguishing between solid-state synthesizable and unsynthesizable compositions (Fig. 8).

The similar performances of model 2 and the supervised learning model highlight the importance of understanding the dataset before model training. Although the supervised learning model appeared to perform better when looking at only the evaluation metrics alone, the model cannot predict the solid-state synthesizability of hypothetical compositions reliably. This is because the supervised learning model learns only from solid-state and non-solid-state synthesized compositions (both are synthesizable compositions), but does not learn from unsynthesizable compositions. Therefore, the supervised learning model can only make solid-state synthesizability predictions on synthesizable compositions, which is not the case for hypothetical compositions, since they contain both synthesizable and unsynthesizable compositions. As a result, the supervised learning model predicted that a high number of hypothetical compositions to be solid-state synthesizable as shown in Table 3.

To further verify the model predictiveness, the predictions made by model 1 on non-solid-state synthesized compositions were examined. An examination of 24 compositions predicted

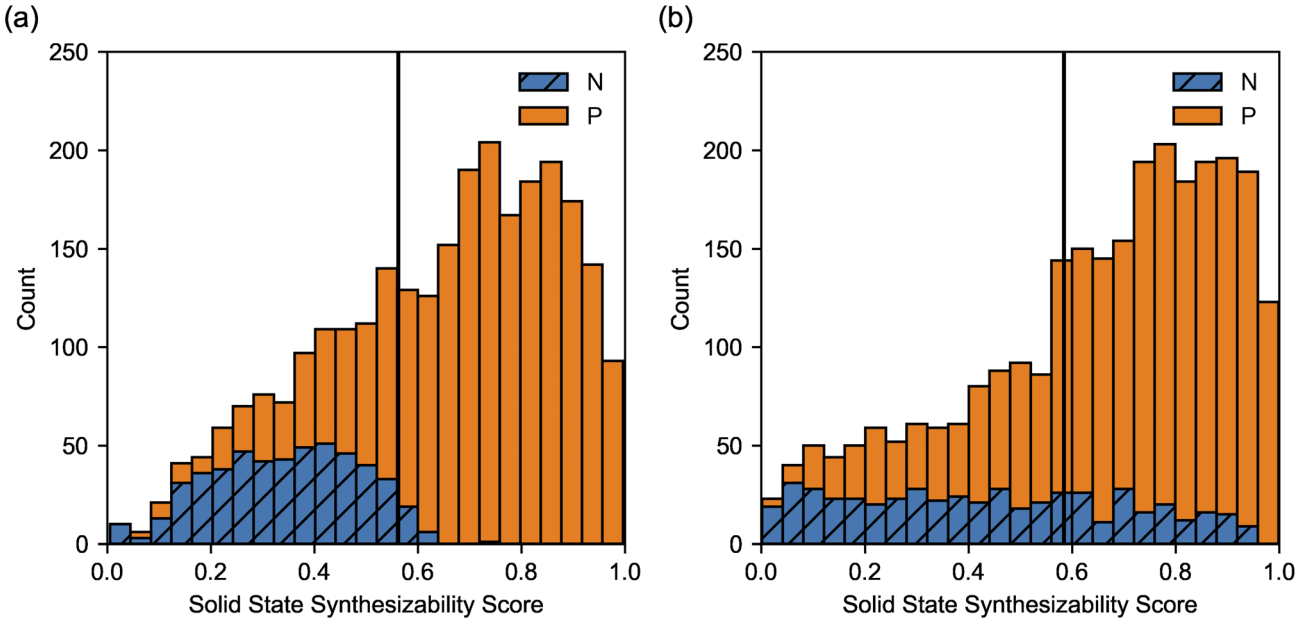

Fig. 7 The solid-state synthesizability score of the synthesized compositions from model 1 (a) and model 2 (b), where the orange bars represent the solid-state synthesized composition, blue stripe bars represent the non-solid-state synthesized composition, and the black lines are the decision threshold. The scores of each entry are the average of the out-of-sample scores during cross-validation.

Table 3 Performances and the percentage of predicted solid-state synthesizable (SSS) hypothetical ternary oxide compositions of model 1, model 2, and the supervised model

Model	Decision threshold	ROC AUC	*G*-mean	TPR	FPR	Predicted SSS (%)
Model 1	0.562	0.921	0.863	0.780	0.047	3.9
Model 2	0.584	0.771	0.713	0.738	0.311	7.9
Supervised model	0.827	0.775	0.767	0.813	0.314	66.3

to be solid-state synthesizable but labeled as non-solid-state synthesized material in the human curated dataset found that 45.8% (11 out of 24) of them have been synthesized *via* solid-state synthesis, which indicates they were erroneous entries in the human curated dataset. A further 40 compositions with the highest score below the decision threshold and the 30 compositions with the lowest scores were examined for comparison, where only 7.5% and 10% were found to have been synthesized *via* solid-state reaction, respectively. The much higher percentage of solid-state synthesized compositions above the decision threshold compared to compositions below the threshold means that model 1 is capable of predicting the solid-state synthesizability of compositions.

3.5.2 Hypothetical compositions. Model 1 predicted that 168 out of 4357 hypothetical compositions (3.9%) are solid-state synthesizable. As a further criterion, model 3 was used to validate the result of model 1, which predicted that 134 out of 168 (79.8%) are synthesizable. The performance of model 3 and its comparison with model 1 are shown in ESI S9 and S10,† respectively. Out of the 134 compositions, at least 56 have been synthesized in the literature, among which at least 43 have been synthesized *via* solid-state reactions. The list of 25 hypothetical ternary oxides that have not been synthesized but are predicted to be solid-state synthesizable with the highest score are shown in Table 4. The full list is available in ESI Table S9.†

During inspection of the candidate compositions, 3 of them were removed as these compositions are highly likely to be synthesized compositions with non-stoichiometric oxygen ($Bi_{16}Ru_{16}O_{55}$ is probably oxygen-deficient $Bi_2Ru_2O_7$) or with fractional occupancies ($Na_{11}(Ru_4O_9)_4$ and $K_6Nb_{11}O_{30}$ are probably $Na_{2.7}Ru_4O_9$ (ref. 57) and $K_6Nb_{10.88}O_{30}$ (ref. 58)). This highlights a limitation of the Materials Project, where disordered materials are not distinctively represented from ordered materials.[59]

3.5.3 Comparison with previous study. A comparison of the general synthesizability scores of the ternary oxide composition in the Materials Project predicted by model 3 and Jang *et al.*[32] is shown in Fig. 9. The overall agreements between the models are 81.7% and 80.5% for the composition with and without ICSD IDs, respectively. The TPR are 0.882 and 0.831 for model 3 and Jang *et al.*, respectively. The major difference between the models is the number of predicted synthesizable compositions for the compositions without ICSD IDs, which are 5.2% for model 3 and 19.3% for Jang *et al.* model.

One possible reason for the difference in prediction is that while both used PU learning, Jang *et al.*'s model used structural

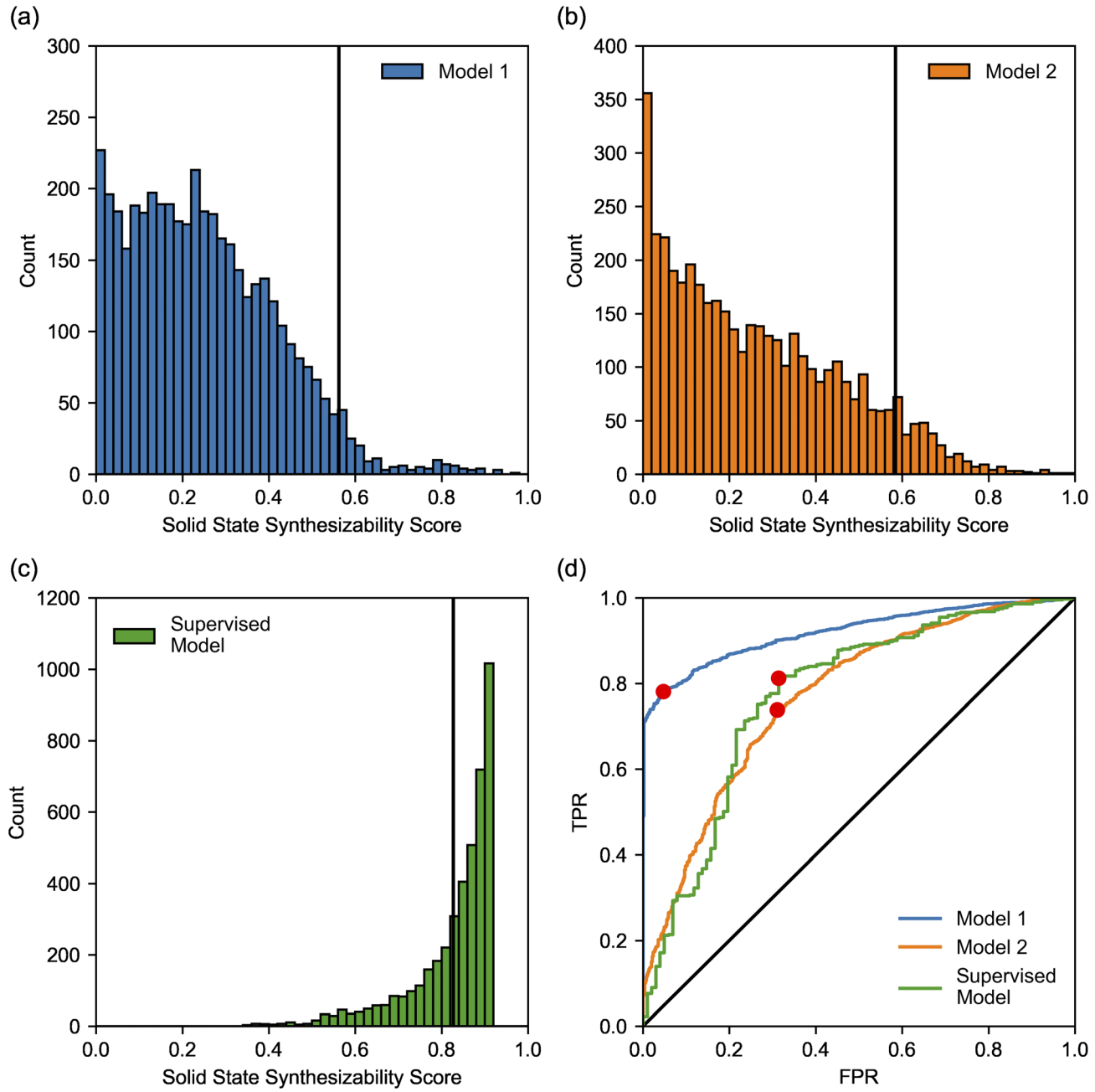

Fig. 8 Solid-state synthesizability score of the hypothetical ternary compositions in Materials Project predicted using (a) model 1, (b) model 2, and (c) supervised learning model. The vertical black lines represent the decision threshold that maximizes *G*-mean. (d) Shows the receiver operating characteristic curves of model 1, model 2, and the supervised learning model based on the scores of the synthesized entries, where the red dots represent the maximized *G*-mean value for each curve and the diagonal black line represents a model that makes random guesses.

features,[32] while model 3 used compositional features. This could mean that a hypothetical material with a dissimilar composition but a similar crystal structure to synthesized materials might be predicted as synthesizable by Jang *et al.*'s model but unsynthesizable by model 3. Another reason is the difference in training data. While model 3 was trained with only ternary metal oxides, Jang *et al.*'s model was trained on materials aside from ternary oxides. In addition, an examination of the Materials Project entries assumed to be synthesized ternary oxides by Jang *et al.* showed that around 10% have a different stoichiometry from the actual material (*e.g.* missing hydrogen for hydrates or oxygen-rich/deficient phases) or is a duplicated entry of the same materials based on a different structure refinement.

3.5.4 Limitation of the method. When applying PU learning, some assumptions needed to be made about the labeling mechanisms (why the data point is labeled or unlabeled) and the class distribution.[29] Often, PU learning assumes that the labeled data are selected completely at random (SCAR) or selected at random (SAR). The former assumes the

Table 4 The solid-state synthesizability score of the 25 ternary oxide compositions that have not been synthesized in the literature with the highest solid-state synthesizability score

Materials project id	Composition	Solid-state synthesizability score
mp-778430	$K_8Al_2O_7$	0.926
mp-1200359	$Na_{12}(CuO_2)_7$	0.895
mvc-3343	$Zn_3Sn_2O_7$	0.869
mp-774805	$Na_6Mn_7O_{10}$	0.861
mp-758139	$Sr_{21}Co_{14}O_{43}$	0.844
mp-1197010	$Sr_5Ti_9O_{23}$	0.841
mp-1223708	$K_2Mo_8O_{13}$	0.838
mp-1096838	$Ba(AgO)_2$	0.827
mp-774428	$K_3V_{14}O_{28}$	0.826
mp-1197629	$Sr_3Ti_5O_{13}$	0.819
mp-1202462	$Sr_5Ti_8O_{21}$	0.812
mp-674312	$Eu_2Nb_4O_{13}$	0.810
mp-1147775	Ba_2OsO_4	0.807
mp-1018032	$SrCdO_2$	0.805
mp-1198567	$Sr_{25}Ti_{39}O_{103}$	0.791
mp-757454	Mn_6PbO_{12}	0.790
mp-1202132	$Sr_5Ti_7O_{19}$	0.789
mp-1201432	$Sr_{15}Ti_{23}O_{61}$	0.787
mp-755102	$K_6Cr_2O_9$	0.781
mp-1208410	$TbMoO_5$	0.746
mp-773070	KNb_2O_5	0.737
mp-753320	RbV_4O_{10}	0.718
mp-981103	Sr_3CdO_4	0.705
mp-674350	$TiPb_9O_{11}$	0.703
mp-1095551	K_8AsO_3	0.691

probability of a positive sample being labeled does not depend on its attributes and that all positive samples have an equal likelihood being labeled.[60] The latter assumes the probability of a positive sample to be labeled depends on its attributes, but not the probability of it being positive.[61] An analogy of SAR is that patients with more severe symptoms are more likely to be identified.[61] In terms of the context of this study, the SCAR assumption would be to assume all synthesizable materials have an equal likelihood of being discovered, while an example of the SAR assumption would be that the probability of a synthesizable material being discovered/labeled depends on how well explored its chemical system is, instead of the probability of it being synthesizable.

The labeling scheme for synthesizable materials is complicated because it depends on the discovery of the materials and the origin of the data. Solid-state synthesizable compositions that are unlabeled in this study can be due to (1) no or failed synthesis attempts; (2) the information is not added to ICSD and the Materials Project; (3) the material has been solid-state synthesized but was not found during data collection. All three reasons are affected by the bias in research on certain chemical systems/crystal structures, either because the materials have desired properties or their relative ease of synthesis. If there is any bias in the discovery of materials and/or inclusion in relevant databases, the model will underestimate the synthesizability of hypothetical materials that are dissimilar to the training data but otherwise synthesizable. Nevertheless, such limitations are not unique to PU learning and apply to other data-driven approaches for materials research. Although biases from the data are not always a negative aspect, *e.g.* exclusion of elements due to toxicity or limited availability, we believe this should be considered when drawing conclusions from the results.

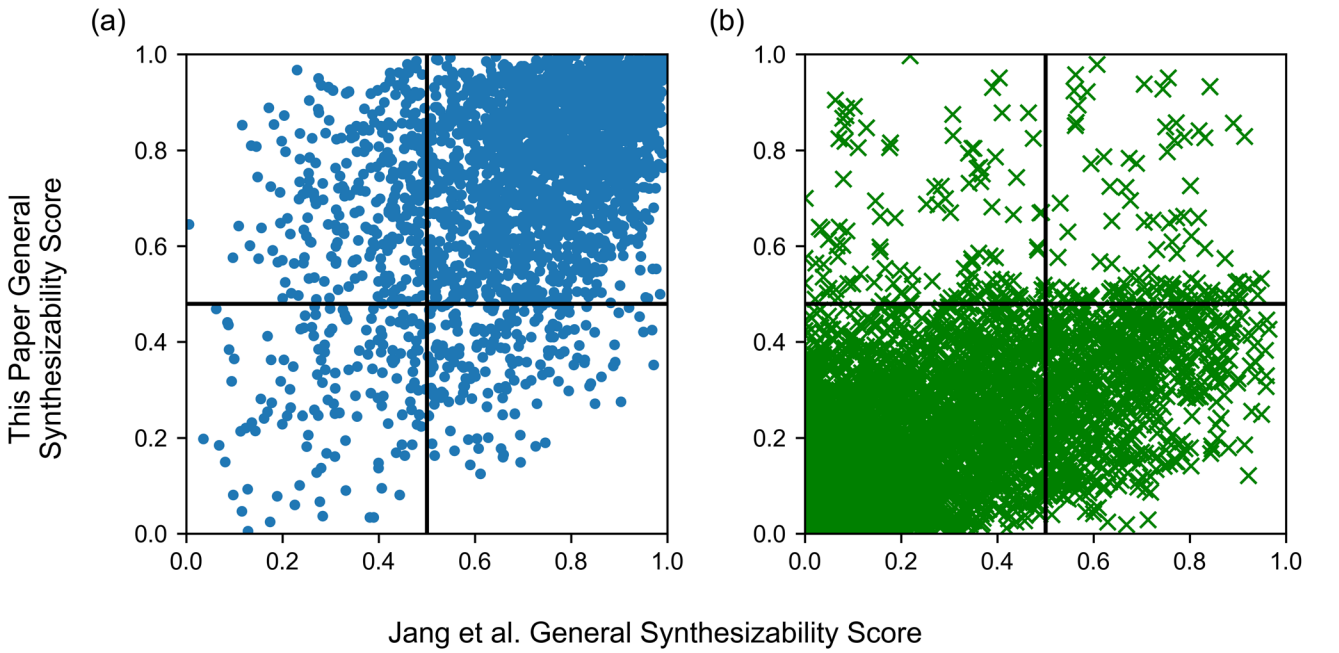

Fig. 9 Comparison of general synthesizability score of the 6924 ternary oxide compositions present in both the current and Jang *et al.* study.[32] (a) 2652 compositions with ICSD IDs (b) 4272 compositions without ICSD IDs. The horizontal and vertical lines are the decision thresholds for model 3 and Jang *et al.* model, respectively. The upper right and lower left quadrants of each plot contain the compositions where the classification of both models align. For compositions with polymorphs in Jang *et al.* study, the highest score of each composition was chosen for this comparison.

The bias in model 1 synthesizability prediction was examined by comparing the chemical systems between the predicted synthesizable and synthesized compositions. 141 out of 168 (83.9%) of predicted solid-state synthesizable compositions are from chemical systems that have at least one solid-state synthesized composition in the human curated dataset. This indicates that a hypothetical composition has a higher likelihood of being predicted as solid-state synthesizable if there are other solid-state synthesized compositions in the training dataset with the same chemical systems.

Another limitation of the PU method in this study is assuming the absence of observation as negative data, where synthesized compositions that haven't been solid-state synthesized are treated as solid-state unsynthesizable. This assumption is supported by the fact that more than 80% of the synthesized compositions in this dataset have been solid-state synthesized in at least one paper, demonstrating that solid-state synthesis has been one of the most popular synthesis approaches. Therefore, it is likely that solid-state synthesis has been attempted for most of the compositions. However, this assumption may not apply to relatively new or uncommon synthesis methods. More importantly, solid-state synthesis is uncommonly chosen for the synthesis of materials where the intended application requires the dimensionality of the materials to be in the nanometer or micrometer scale. Therefore, exploratory investigations to discover new thin-film or nanomaterials might be biased toward synthesis methods like sol–gel and thin-film deposition. In these cases, the evaluation of the false positives would be difficult and might lead to an overestimation or underestimation of the number of synthesizable materials.

4 Conclusions

In this work, a dataset on the solid-state synthesis of ternary metal oxides was human curated with the intended goal of supplementing text-mined datasets and synthesis planning. The dependence of synthesizability on the synthesis method was explored, and the analysis of the human curated dataset showed that the calculated thermodynamic convex hull could not distinguish between solid-state synthesized and non-solid-state synthesized compositions. Furthermore, the trends from the analysis were applied to identify outliers in text-mined datasets containing solid-state reaction information. Examination of outliers reveals how small details in the data annotation stage of text-mining may result in the omission of data and erroneous extracted information, which would deteriorate the predictive power of any derived models.

This human curated dataset was then used to train a PU learning model to predict the solid-state synthesizability of hypothetical ternary oxide compositions using only compositional information. By cross-validating with another model that predicts the general synthesizability, 134 compositions were identified to be solid-state synthesizable, of which at least 56 have been synthesized and 43 out of the 56 have been solid-state synthesized. Future investigations on synthesizing hypothetical compositions can be attempted by predicting synthesis conditions using the collected data.

Data availability

The code and data used in this study (including evaluation of the dataset) can be found at **https://doi.org/10.5281/zenodo.7726604**.

Author contributions

A. W. and D. P. supervised the project. All the authors conceptualized the project. V. C. curated the dataset, developed the software, and analyzed the data. All the authors discussed the results and co-wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1 and EP/T022213/1). We also wish to acknowledge the use of the EPSRC funded Physical Sciences Data-science Service hosted by the University of Southampton and STFC under grant number EP/S020357/1.

References

1 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
2 G. Bergerhoff, R. Hundt, R. Sievers and I. D. Brown, *J. Chem. Inf. Comput. Sci.*, 1983, **23**, 66–69.
3 K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N. G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S. H. Wei and J. Perkins, *The 2019 Materials by Design Roadmap*, 2019, vol. 52.
4 C. J. Bartel, A. W. Weimer, S. Lany, C. B. Musgrave and A. M. Holder, *npj Comput. Mater.*, 2019, **5**, 4.
5 J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti and M. A. Marques, *Chem. Mater.*, 2017, **29**, 5090–5103.
6 H. Liu, J. Cheng, H. Dong, J. Feng, B. Pang, Z. Tian, S. Ma, F. Xia, C. Zhang and L. Dong, *Comput. Mater. Sci.*, 2020, **177**, 109614.
7 A. K. Singh, J. H. Montoya, J. M. Gregoire and K. A. Persson, *Nat. Commun.*, 2019, **10**, 443.
8 W. Li, R. Jacobs and D. Morgan, *Comput. Mater. Sci.*, 2018, **150**, 454–463.

9 S. Kirklin, J. E. Saal, V. I. Hegde and C. Wolverton, *Acta Mater.*, 2016, **102**, 125–135.
10 M. Aykol, S. S. Dwaraknath, W. Sun and K. A. Persson, *Sci. Adv.*, 2018, **4**, eaaq0148.
11 C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanović, C. B. Musgrave and A. M. Holder, *Nat. Commun.*, 2018, **9**, 4168.
12 B. R. Chen, W. Sun, D. A. Kitchaev, J. S. Mangum, V. Thampy, L. M. Garten, D. S. Ginley, B. P. Gorman, K. H. Stone, G. Ceder, M. F. Toney and L. T. Schelhas, *Nat. Commun.*, 2018, **9**, 1–20.
13 P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
14 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 1–11.
15 C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M. Scheffler, *Sci. Adv.*, 2019, **5**, 1–10.
16 E. Kim, K. Huang, A. Saunders, A. Mccallum, G. Ceder and E. Olivetti, *Chem. Mater.*, 2017, **29**, 9436–9444.
17 E. Kim, E. Strubell, A. Tomala, A. McCallum, S. Matthews, E. Olivetti, A. Saunders and K. Huang, *Sci. Data*, 2017, **4**, 170127.
18 E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H. S. Chang, E. Strubell, A. McCallum, S. Jegelka and E. Olivetti, *J. Chem. Inf. Model.*, 2020, **60**, 1194–1201.
19 O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan and G. Ceder, *Sci. Data*, 2019, **6**, 203.
20 S. A. Malik, R. E. A. Goodall and A. A. Lee, *Chem. Mater.*, 2021, **33**, 616–624.
21 H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain and G. Ceder, *Chem. Mater.*, 2022, **34**, 7323–7336.
22 C. Karpovich, E. Pan, Z. Jensen and E. Olivetti, *Chem. Mater.*, 2023, **35**(3), 1062–1079.
23 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, *Nature*, 2023, **624**, 86–91.
24 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.
25 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
26 A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens and T. Laino, *Nat. Mach. Intell.*, 2021, **3**, 485–494.
27 O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti and G. Ceder, *iScience*, 2021, **24**, 1–20.
28 Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma and E. Olivetti, *ACS Cent. Sci.*, 2019, **5**, 892–899.
29 J. Bekker and J. Davis, *Mach. Learn.*, 2020, **109**, 719–760.
30 F. Mordelet and J.-P. Vert, *Pattern Recognit. Lett.*, 2014, **37**, 201–209.
31 N. C. Frey, J. Wang, G. I. Vega Bellido, B. Anasori, Y. Gogotsi and V. B. Shenoy, *ACS Nano*, 2019, **13**, 3031–3041.
32 J. Jang, G. H. Gu, J. Noh, J. Kim and Y. Jung, *J. Am. Chem. Soc.*, 2020, 1–15.
33 G. H. Gu, J. Jang, J. Noh, A. Walsh and Y. Jung, *npj Comput. Mater.*, 2022, **8**, 71.
34 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
35 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, *npj Comput. Mater.*, 2019, **5**, 62.
36 *Handbook of Chemistry and Physics*, ed. J. R. Rumble, CRC Press, Taylor & Francis Group, 2020, 101st edn, **https://hbcp.chemnetbase.com/faces/contents/ContentsSearch.xhtml**.
37 V. Raghavan, P. Bollmann and G. S. Jung, *ACM Trans. Inf. Syst.*, 1989, **7**, 205–229.
38 L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, *Comput. Mater. Sci.*, 2018, **152**, 60–69.
39 Welcome to LightGBM's Documentation! — LightGBM 3.2.1.99 Documentation, **https://lightgbm.readthedocs.io/en/latest/index.html**.
40 S. Jain, M. White and P. Radivojac, *Recovering True Classifier Performance in Positive-Unlabeled Learning*, 2017.
41 E. Briz-López, M. Ramírez-Moreno, I. Romero-Ibarra, C. Gómez-Yáñez, H. Pfeiffer and J. Ortiz-Landeros, *J. Energy Chem.*, 2016, **25**, 754–760.
42 R. S. Barker and I. R. Evans, *J. Solid State Chem.*, 2006, **179**, 1918–1923.
43 S. R. Young, A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin and R. K. Vasudevan, *J. Appl. Phys.*, 2018, **123**, 115303.
44 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, *Sci. Adv.*, 2016, **2**, e1600225.
45 C. Toher, C. Oses, D. Hicks and S. Curtarolo, *npj Comput. Mater.*, 2019, **5**, 10–12.
46 C. M. Rost, E. Sachet, T. Borman, A. Moballegh, E. C. Dickey, D. Hou, J. L. Jones, S. Curtarolo and J. P. Maria, *Nat. Commun.*, 2015, **6**, 8485.
47 P. K. Todd, A. Wustrow, R. D. McAuliffe, M. J. McDermott, G. T. Tran, B. C. McBride, E. D. Boeding, D. O'Nolan, C.-H. Liu, S. S. Dwaraknath, K. W. Chapman, S. J. L. Billinge, K. A. Persson, A. Huq, G. M. Veith and J. R. Neilson, *Inorg. Chem.*, 2020, **59**, 13639–13650.
48 C. M. Fernández-Posada, A. Castro, J. M. Kiat, F. Porcher, O. Peña, R. Jiménez, M. Algueró and H. Amorín, *Adv. Funct. Mater.*, 2018, **28**, 1–11.
49 T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari and G. Ceder, *Chem. Mater.*, 2020, **32**, 7861–7873.
50 F. Zhang, M. Zhu and C. Wang, *Int. J. Refract. Metals Hard Mater.*, 2008, **26**, 329–333.
51 D. Zhang, R. Cai, Y. Zhou, Z. Shao, X.-Z. Liao and Z.-F. Ma, *Electrochim. Acta*, 2010, **55**, 2653–2661.

52 Y. Hashimoto, M. Wakeshima and Y. Hinatsu, *J. Solid State Chem.*, 2003, **176**, 266–272.
53 A. A. Belik, T. Wuernisha, T. Kamiyama, K. Mori, M. Maie, T. Nagai, Y. Matsui and E. Takayama-Muromachi, *Chem. Mater.*, 2006, **18**, 133–139.
54 Z. Chen, H. Zhu, S. Ji, V. Linkov, J. Zhang and W. Zhu, *J. Power Sources*, 2009, **189**, 507–510.
55 R. Merkle and J. Maier, *Z. Anorg. Allg. Chem.*, 2005, **631**, 1163–1166.
56 D. Jha, L. Ward, A. Paul, W. keng Liao, A. Choudhary, C. Wolverton and A. Agrawal, *Sci. Rep.*, 2018, **8**, 1–13.
57 K. Regan, Q. Huang, M. Lee, A. Ramirez and R. Cava, *J. Solid State Chem.*, 2006, **179**, 195–204.
58 P. Becker and P. Held, *Z. Kristallogr. New Cryst. Struct.*, 2000, **215**, 319–320.
59 J. Leeman, Y. Liu, J. Stiles, S. B. Lee, P. Bhatt, L. M. Schoop and R. G. Palgrave, *PRX Energy*, 2024, **3**, 011002.
60 C. Elkan and K. Noto, *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD 08*, Las Vegas, Nevada, USA, 2008, p. 213.
61 J. Bekker, P. Robberechts and J. Davis, in *Machine Learning and Knowledge Discovery in Databases*, ed. U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis and C. Robardet, Springer International Publishing, Cham, 2020, vol. 11907, pp. 71–85.