MACHINE LEARNING
Science and Technology

PURPOSE-LED
PUBLISHING™

Impact of data bias on machine learning for crystal compound synthesizability predictions

To cite this article: Ali Davariashtiyani *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 040501

View the article online for updates and enhancements.

ECS The Electrochemical Society
Advancing solid state & electrochemical science & technology
247th ECS Meeting
Montréal, Canada
May 18-22, 2025
Palais des Congrès de Montréal
Showcase your science!
Abstract submission deadline extended: December 20
ECS UNITED

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
23 March 2024

REVISED
16 September 2024

ACCEPTED FOR PUBLICATION
15 November 2024

PUBLISHED
26 November 2024

BENCHMARK

Impact of data bias on machine learning for crystal compound synthesizability predictions

Ali Davariashtiyani[1], Busheng Wang[2], Samad Hajinazar[2], Eva Zurek[2] and Sara Kadkhodaei[1,*]

[1] Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago, Chicago, IL, United States of America
[2] Department of Chemistry, State University of New York at Buffalo, Buffalo, NY, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: sarakad@uic.edu

Keywords: machine learning, data bias, synthesizability, crystal compounds

Abstract

Machine learning models are susceptible to being misled by biases in training data that emphasize incidental correlations over the intended learning task. In this study, we demonstrate the impact of data bias on the performance of a machine learning model designed to predict the likelihood of synthesizability of crystal compounds. The model performs a binary classification on labeled crystal samples. Despite using the same architecture for the machine learning model, we showcase how the model's learning and prediction behavior differs once trained on distinct data. We use two data sets for illustration: a mixed-source data set that integrates experimental and computational crystal samples and a single-source data set consisting of data exclusively from one computational database. We present simple procedures to detect data bias and to evaluate its effect on the model's performance and generalization. This study reveals how inconsistent, unbalanced data can propagate bias, undermining real-world applicability even for advanced machine learning techniques.

1. Introduction

The increasing availability of large materials datasets has facilitated the rapid growth of machine learning for materials discovery and design [1–15]. However, the impact of data selection strategies, data quality, or data bias has been less explored or not clearly demonstrated, even for the most advanced machine learning models, although these factors can significantly impact model performance, generalizability, and reliability [16, 17]. In this study, we specifically investigate the effect of *data heterogeneity* on machine learning models. For demonstration, we use a machine learning model designed for predicting crystalline compounds' synthesis feasibility (or synthesizability), previously developed by our group [18]. Our machine learning model for synthesizability prediction is an ideal platform because data selection for such a model is inherently challenging. A portion of the data must represent already synthesized materials (synthesizable) from crystal structure datasets, while another portion must be artificially generated to represent hypothetical crystals unlikely to be synthesized or formed (unsynthesizable). This challenge has led to arbitrary data selection in the literature, resulting in different strategies. For example, studies in [19–22] assume that unsynthesizable examples are not available and thus predicts 2D or crystalline materials' synthesizability based on a positive and unlabeled (PU) classification model [23] on data from a single dataset. In contrast, [18] predicts crystalline materials' synthesizability based on a binary classifier on labeled data from different sources. Here, we examine the generalization performance of our synthesizability model for two data selection approaches: whether data is coming from different sources (heterogeneous data) or from an identical source (homogeneous data).

Inherent data bias is a common challenge in machine learning [17], stemming from sampling bias, data collection bias, domain bias, or labeling bias. The bias in data usually leads to spurious correlations and biases picked up by the model, regardless of the choice of the machine learning algorithm [17]. In the field of data-driven materials research, some studies have explored the influence of data on the performance of machine learning models [24–28]. For example, Kumagai *et al* demonstrated that data bias influences the error and reliability of predictions made by a machine learning model [24]. Using the experimental property data from the Starrydata2 database as a demonstration platform, they defined an applicability domain—a material space to which the machine learning model can be applied based on its similarity to known materials used for training. They revealed that model predictions are more reliable within the applicability domain. Within the applicability domain, prediction accuracy remains high, while outside the defined applicability domain, accuracy significantly decreases. They also observed that the prediction error decreases within the applicability domain as the number of neighboring known materials increases. In another study, Zhang *et al* [25] introduced an information entropy-based metric to measure data bias and developed an entropy-targeted active learning (ET-AL) framework to mitigate it. They utilized the ET-AL framework to guide new data acquisition in density functional theory (DFT)-generated materials databases (such as OQMD [29, 30] and JARVIS [31]) by addressing the imbalanced coverage of formation energy among different crystal systems (i.e. structure-stability bias). This approach aimed to enhance the diversity of underrepresented crystal systems and, as a result, improved the performance of machine learning models [25]. In a separate study, Li *et al* [26] investigated the impact of distribution shifts (or domain shifts) on the performance of machine learning models, offering solutions for diagnosing, anticipating, and addressing this challenge. Data distribution can undergo significant shifts, even between different versions of an actively expanding database, owing to changes in preferences or focus over time. For instance, they illustrated a significant decline in the performance of a formation energy prediction model trained on Materials Project 2018 (e.g. ALIGNN-MP18 [32]) when applied to new compounds in the Materials Project 2021 database [33], with prediction errors ranging from 23 to 160 times larger than those observed when the model is tested on Materials Project 2018. By utilizing a uniform manifold approximation and projection (UMAP) approach as a measure of similarity to cluster the data, they observed that test samples with low prediction errors tend to reside within clusters covered by the training data (i.e. similar to the training data). Conversely, the majority of poorly predicted test samples form a distinct, isolated cluster separate from the rest of the data. To enhance prediction robustness and generalizability of the machine learning model, they introduced both UMAP-guided and query-by-committee data acquisition strategies. Other studies noted the impact of data bias on machine learning model performance [34, 35]. For example, our group observed that the bias in the DFT materials database, specifically the imbalanced distribution of negative and positive formation energies in the Materials Project, results in diminished prediction performance for larger, positive-value ranges of formation energy and is the likely cause for the progressive increase of error from metallic to ionic materials [34].

As a demonstration platform for exploring data bias, we utilize a machine learning model developed by our group to predict the synthesizability of crystal compounds [18], albeit with significant modifications that enhance its performance (as detailed in the Methods section). Therefore, we present a concise review of machine-learning methods applied to materials synthesis, categorizing these studies into two groups: the first group are studies that focus on developing machine learning models for predicting the synthesis feasibility (i.e. synthesizability) of given products or crystal compounds [18–22, 36, 37]. The demonstrative model used in this study belongs to this group. These models typically involve learning correlations between chemical/structural patterns in existing crystal compounds and a score of synthesizability (or synthesizability likelihood). As mentioned earlier, data selection and labeling are difficult in such models due to the absence of unsynthesizable crystal compounds. The second group of studies aims to develop models for predicting synthesis routes or reactions (e.g. solid-state, sol-gel, or solution—hydrothermal, precipitation), synthesis procedures, synthesis conditions (e.g. temperatures, times), or synthesis precursors or reactants [38–47]. These studies encompass a range of approaches, from data-driven learning of materials synthesis information using natural language processing of existing scientific literature [39–45], to the development of graph-based networks based on thermodynamic and kinetic data (i.e. physics-informed) [47–50]. The latter approach is employed for predicting chemical reaction pathways in solid-state materials synthesis.

In this study, we analyze the performance of a crystal compound synthesizability model trained on two sets of data through a comparative modeling experiment. In the first experiment, data is collected from two distinct sources, creating a mixed-source or heterogeneous dataset. Synthesizable crystal compounds are sourced from the Crystallography Open Database (COD) [51], while unsynthesizable samples are computationally generated using the crystal structure prototype database (CSPD) [52]. Details of the generation of unsynthesizable samples are provided in the Methods section and in appendix C. In the second experiment, data for both classes is collected from a single dataset, specifically the DFT-generated Materials Project database [33]. We refer to this dataset as a single-source or homogeneous dataset. The atomic configurations of the reported crystal compounds in Materials Project, compromising the single-source dataset, are relaxed through DFT geometry and cell optimizations. Therefore, the crystal structure data from the Materials Project corresponds to zero-pressure and no-applied-stress conditions within the DFT calculations. For the mixed-source dataset, the COD dataset does not report pressure for every entry, but most samples were likely obtained under ambient conditions (details below). In our data collection approach pressure was not specified nor was inferred for the mixed dataset. Appendix C provides details for CSPD structure generation. In both experiments, we employ the same machine learning architecture—a convolutional neural network (CNN) for processing sparse voxel image representations of crystals connected to a binary classifier. Further details of the machine learning model are provided in the Methods section. We compare the predictive performance and generalizability of both models and investigate the underlying reasons for performance differences.

2. Methods

Crystal synthesizability model

Our crystal synthesizability model consists of a CNN connected to a neural network classifier. The interconnect CNN and neural network perform feature learning and classification tasks, respectively, on labeled crystal structure data. This model operates on sparse voxel images of crystals, as developed by our group [18, 34]. Our voxel image representation creates a 3D visual depiction of the crystal structure color-coded by the identities of its constituent chemical elements (see [34] for details). The voxel images are created in a cubic box with a 50 Å edge. The atoms in the crystal unit cell are repeated to fill the cube. Then, the cube is partitioned into a voxel grid of size $128 \times 128 \times 128$. The voxel images are RGB color-coded with the three channels representing atomic number, group number, and valence number, respectively. The CNN encodes the hidden structural and chemical patterns of crystal compounds by processing the sparse voxel images into a low-dimension set of latent features. The latent features are input into the neural network classifier (a binary classifier). The model is trained on labeled crystal images, representing two classes of synthesizable and unsynthesizable crystals. The architecture of the synthesizability model is shown in figure 1 and is detailed in our previous work [18]. The CNN consists of a sequence of three blocks, each consisting of a convolution, an activation, and a pooling layer. The CNN architecture flattens the output and propels it through a 3-layer dense neural network with a (13,13,13,1) node architecture. This model is operated on our most recent framework for sparse voxel image representation of crystalline materials as detailed in [34]. Compared to our previous study [18], the synthesizability model used in this work adopts a more advanced voxel image representation (as detailed in [34]). Additionally, the presented synthesizability model employs augmentation of rotated crystal samples (data augmentation) during training and an ensemble averaging technique during prediction to improve the model's consistency and rotational invariance. Details of the data augmentation and ensemble averaging are provided in [34] with a brief discussion in appendix A.

Labeling approach

The binary classification is performed on two classes of crystalline materials: synthesizable versus unsynthesizable crystals, the latter being the hypothetical crystalline materials that are unlikely to be synthesized. The positive or synthesizable class comprises experimentally synthesized crystal compounds readily available in crystal databases (e.g. inorganic crystal structure database (ICSD) [53] and COD [51]). On the other hand, the negative or unsynthesizable class must represent crystal compounds that are unlikely to form or be experimentally synthesized, at least based on the existing scope of synthesis techniques and conditions available to us. Therefore it lacks a dedicated repository, which makes the collection of data for this class challenging. Other models for synthesizability in previous studies [19, 20, 22] use PU learning as they assume the absence of explicit negative class samples. In contrast, we use an *a priori* labeling approach by carefully selecting negative class samples. Our strategy involves identifying the top 0.1% of well-studied crystal compounds in the materials science literature from 1922 to 2021, resulting in 108 compositions or chemical formulas, as presented in appendix B. The rationale behind this approach is that these compositions have been extensively explored, ensuring that all possible synthesizable crystal structures have likely been realized. Those hypothetical polymorphs of these compositions that have not been synthesized are most likely unsynthesizable. We employed a natural language processing tool to select these compositions, as detailed in [18]. For these compositions, we assign a negative label to crystal structure polymorphs not found in existing experimental databases. Crystal samples come from either a single source or multiple sources, as explained below.

In contrast to our *a priori* labeling approach, alternative synthesizability models employ a semi-supervised learning strategy, incorporating both labeled and unlabeled data (PU learning). This involves learning characteristics of negative samples through a data-driven machine learning technique known as pseudo-labeling. Notable examples include the transductive bagging scheme utilized in [19–21], and the dynamic entropy-based pseudo-labeling within a teacher–student dual neural network [22]. In PU learning, the model is trained to learn characteristics associated with positive samples by distinguishing them

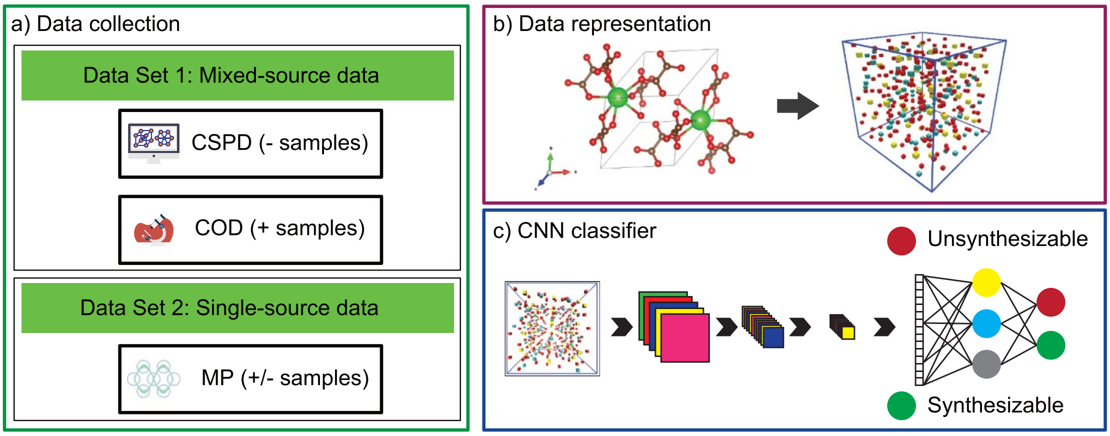

Figure 1. The overall framework of the synthesizability likelihood prediction model. (a) Data collection for the two distinct data sets used in this study. For data set 1 or mixed-source data, crystal samples for the synthesizable class are obtained from the Crystallographic Open Database (COD) and crystal samples for the synthesizable class are generated using the CSPD. For data set 2 or single-source data, crystal samples for both classes are collected from the Materials Project database. (b) The crystal information files (CIFs) are converted into color-coded voxel images, which are used as the inputs for the convolutional encoder. (c) The convolution encoder followed by a multi-layer perceptron (MLP) binary classifier trained on labeld data, referred to as the CNN classifier.

from the 'average' characteristics of unlabeled data, according to a similarity or distance measure between unlabeled and positive samples. In contrast, our synthesizability model learns distinguishing characteristics of positive and negative samples from explicit examples of each class. Therefore, it can be considered a more supervised approach that incorporates a level of human-based physical interpretation of the negative samples.

Both the semi-supervised learning approach and the explicit labeling approach introduce biases into the data and subsequent learning processes (i.e, labeling bias). In our labeling approach in this work, inherent bias arises due to the limited chemical distribution of explicit negative samples, encompassing only 108 chemical compositions as examples of the negative class. We opt against expanding the chemical distribution of negative samples by selecting a larger percentile of top-studied compositions in the literature. This choice is motivated by the need to maintain confidence in designating negative examples, avoiding potential mislabeling. On the other hand, the PU learning approach may introduce other biases. For instance, any unlabeled sample representing a hypothetical crystal compound can contribute to patterns correlated with a negative sample through a weighted average scheme based on similarity measures. However, many of these unlabeled samples collected from computational databases are examples of undiscovered or unexplored synthesizable polymorphs. Therefore, the PU approach has the potential to introduce implicit mislabeling biases to the learning process. Identifying and addressing implicit biases associated with negative samples are beyond the scope of this study. Instead, our focus is on examining the biases introduced through the collection of examples for the positive and negative classes, whether from a single database or multiple databases.

Mixed-source data collection

The mixed-source dataset is compiled from two distinct databases. Positive examples, representing synthesizable materials, are sourced from the experimental COD [51, 54–58], while negative samples, indicating unsynthesizable materials, are generated using the CSPD[52]. The procedure for generating negative samples is as follows: initially, we select the topmost studied compositions in the literature, resulting in 108 unique compositions (details provided in appendix B). These chemical compositions are then input into the CSPD toolkit to generate all possible hypothetical crystal polymorphs. For a given chemical formula, CSPD selects known crystal prototypes from its database as templates. In this process, elemental sites are substituted with the desired chemical elements, and the lattice parameters are adjusted to match a target volume. Finally, the inter-atomic distances of the generated structures undergo validation. Details of utilizing CSPD for generating crystal samples is given in appendix C. From the structures generated by CSPD, we filter out those having identical crystal structures in COD, designating them as belonging to the positive class. Following this procedure, we generated 597 negative samples. For the positive class, we limit the collection to 2960 crystal samples from COD to maintain a balanced sample size between positive and negative classes. The positive samples encompass all crystal polymorphs for the selected 108 chemical compositions available in COD (367 crystals), with the remaining samples randomly chosen from other compositions (2633 crystals). Some of these crystals are filtered out due to image resolution constraints (see [18] for details), resulting in 2960 positive crystal samples passable to the model. The data including positive and negative samples is split into training (60%), validation (20%), and test (20%) sets.

Single-source data collection

The single-source dataset is exclusively compiled from crystal structures in Materials Project (MP v2022.10.28) [33], which is a DFT database for crystal compounds. To ensure consistency between the two datasets for the comparative analysis in this study, we reference the same chemical compositions constituting the mixed-source dataset. For the compositions in the positive class, any MP crystal structure entry that matches the composition and has an ICSD tag is collected as a positive sample. MP entries that match the mixed-source dataset's negative class compositions without an ICSD tag are identified as negative samples. Following this approach, a total of 1068 positive and 930 negative crystal structures are collected from MP. We split the data into training (60%), validation (20%), and test (20%) sets.

3. Results

3.1. Data bias detection

To identify bias within a dataset, one needs to define metrics for measuring such bias. In this study, we employ density and mean atomic mass as two metrics to detect potential biases in both single-source and mixed-source datasets. These features, chosen to represent fundamental characteristics of crystal samples, are selected with the awareness that they are not expected to exhibit a direct correlation with synthesizability. Noticeable differences in the distributions of these basic features between the two classes indicate the presence of data bias that could adversely impact model performance. Distribution discrepancies across datasets have been used in other studies to reveal potential data bias. For instance, Kumagai *et al* [24] compared the distributions of average atomic masses and average electronegativities across different databases, including ICSD, materials project, magnetic materials, and Starrydata2. The observed differences in distributions were considered as a potential source of data bias.

Figures 2(a) and (b) compares the density distributions between synthesizable and unsynthesizable samples for both the single-source and mixed-source datasets. Density is calculated as the total atomic mass divided by the crystal structure volume. As shown in figure 2(a), the density distribution of synthesizable samples is significantly different from the unsynthesizable class in the mixed-source data. Unsynthesizable samples demonstrate a substantially wider density range with a flat peak and a higher average density. Synthesizable samples exhibit a relatively sharp peak around 2 g/cc, while unsynthesizable sample densities widely spread from 2 to 15 g/cc. The higher densities observed in unsynthesizable samples are likely a result of the low-fidelity procedure used to generate hypothetical crystal samples in CSPD, especially when assigning a target volume to the hypothetical crystal. This contrasts with a more accurate albeit computationally intense approach, such as relaxing the crystal volume based on DFT forces (as is done for any hypothetical crystal in the MP database). CSPD constructs a new chemical arrangement of atoms on a crystal prototype skeleton based on simple norms (e.g. composition, symmetry, and configuration) and similarity to available crystal prototypes in its database. As a result of this high-throughput procedure, the target volumes assigned by CSPD to the unsynthesizable samples are systematically smaller than the equilibrium (or relaxed) volume, as evident in figure 2(e). In contrast, the single-source data exhibit similar atomic volume distributions between the negative and positive samples, as shown in figure 2(f). It is important to note that the majority of experimentally synthesized crystals are associated with ambient pressure conditions. For example, our query on the 2023 version of the ICSD database shows that 208 954 crystal entries are associated with pressures below 1 MPa (i.e. ambient pressure) while only 748 crystal entries are for pressure above 1 MPa. The distribution of ICSD crystal samples under pressure (above 1 MPa) is shown in figure E.1 in appendix E. While the pressure information in the ICSD signifies the external conditions during synthesis, it is reasonable to assume that the resulting synthesized crystal reaches equilibrium with its external environment and thus corresponds to the external pressure condition. The systematically larger density of the unsynthesizable samples compared to synthesizable ones indicates a source of bias in the mixed-source data (see figure 2(e)). The illustration of data bias in figure 2(a) and previous studies [24] underscores the need for caution when collecting data from distinct databases, as is the case in this study with mixed-source data. Such data become susceptible to inherent biases measured based on different basic characteristics. The single-source data indicates similar density distributions between the synthesizable and unsynthesizable samples, as shown in figure 2(b). The density distributions for positive and negative samples are both centered around 3 g/cc, although the positive class shows a bimodal

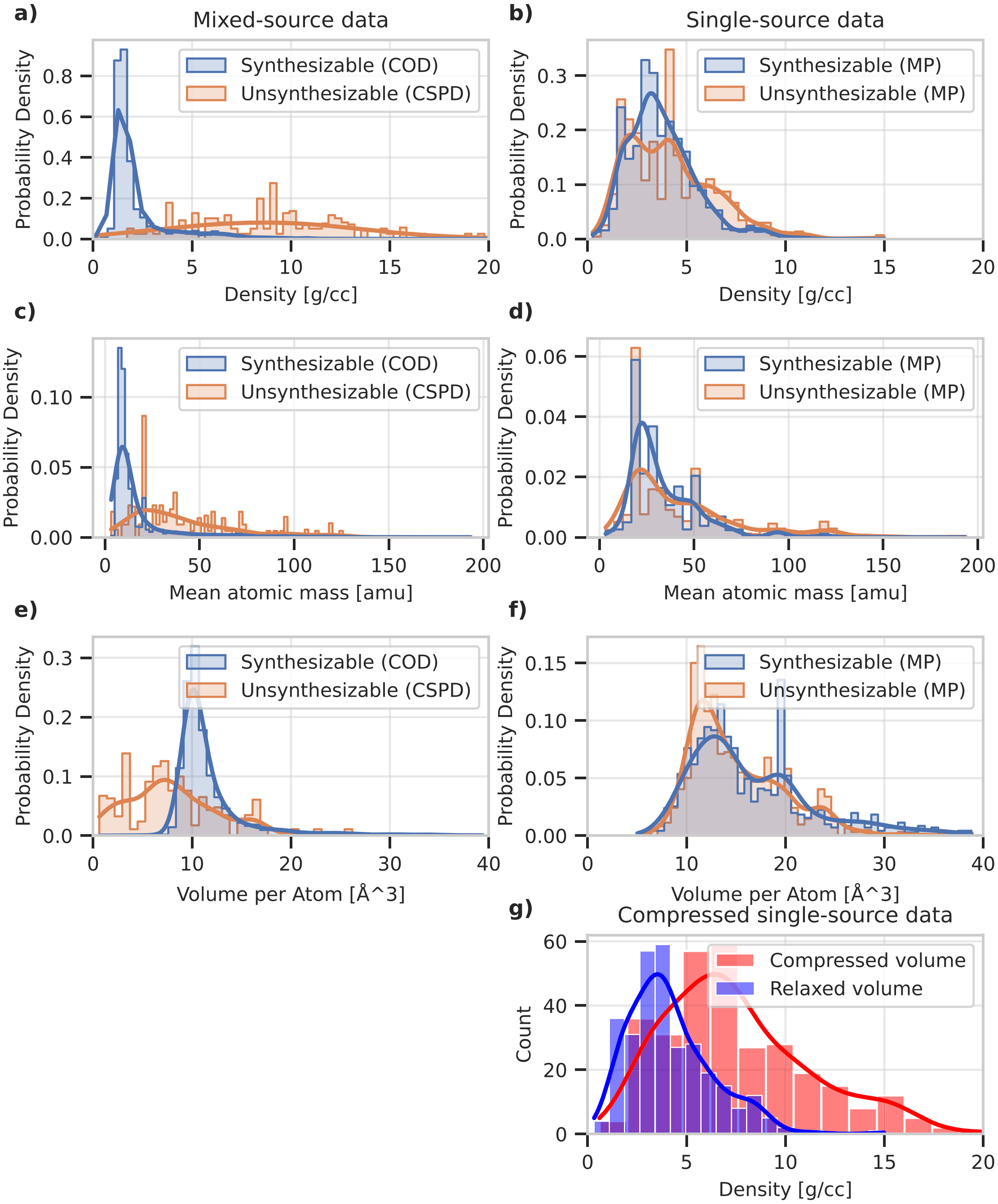

Figure 2. Density, mean atomic mass, and atomic volume distributions of crystal samples across different classes and datasets. (a), (b) Density, (c), (d) mean atomic mass, and (e), (f) atomic volume distributions of crystal samples (synthesizable versus unsynthesizable) in the mixed-source and single source datasets, respectively. (g) Density distributions of crystal samples of the single-source test set, shown for their original (relaxed) volume and compressed volumes by 55%. Both synthesizable and unsynthesizable samples are included in the test set.

distribution with peaks slightly below and above the center. Unlike the mixed-source data, the single-source data does not exhibit any systematic density shift between the synthesizable and unsynthesizable samples.

Apart from density, we compare the mean atomic mass distributions of the two classes in the single-source and mixed-source datasets, as shown in figures 2(c) and (d). We observe noticeable differences in the mean atomic mass distributions between the synthesizable and unsynthesizable classes in the mixed-source data. However, this difference is not as significant as the density distributions. As illustrated in figure 2(c), in the mixed-source data, the synthesizable samples cluster around lower mean atomic masses compared to the unsynthesizable ones. In contrast, the mean atomic mass is more evenly distributed between the synthesizable and unsynthesizable samples in the single-source data (see figure 2(d)). The discrepancy in mean atomic mass distributions in the mixed-source data likely arises from the more diverse chemical compositions accessible in the CSPD database compared to the MP database. In both the single-source and mixed-source data, we draw unsynthesizable samples from the exact same 108 unique chemical formulas. However, the high-throughput nature of the CSPD approach for generating crystal structures results in a more diverse set of crystal structures covering most of the 108 compositions, thereby exhibiting a diverse distribution among different mean atomic masses. On the other hand, for a computationally expensive DFT database such as MP, crystal samples associated with these 108 chemical formulas are less diverse, leading to a sharper peak in the mean atomic mass distribution for negative samples in the single-source data (see figure 2(d)). Comparing figures 2(c) and (d) indicates a more balanced coverage of mean atomic mass between the positive and negative classes in the single-source data, while the negative class shows a larger average mean atomic mass in the mixed-source data. The distribution of other features can be examined between classes to detect potential biases. For example, as shown in appendix D, the distribution of atomic species does not exhibit any significant bias in either the mixed-source or single-source datasets. The observed density distribution discrepancy or imbalance is a result of bias introduced in the mixed-source data rather than any realistic correlation between synthesizability and mean atomic mass.

The baseline feature comparisons in this section reveal clear evidence of data bias between the synthesizable and unsynthesizable samples in the mixed-source data. The balanced distribution of density and mean atomic mass between the two classes in the single-source data highlights the benefits of compiling data from a single source to minimize potential biases. If data collection from distinct sources is necessary, then practices should be considered to make the data from different sources consistent or mitigate the bias they introduce into the model. In the next section, we demonstrate how our machine learning model is susceptible to learning the incidental associations introduced by the bias between classes rather than capturing the intended structural and chemical signatures of synthesizability likelihood.

3.2. Impact of data bias on model performance

Learning performance

Figures 3(a) and (b) compares the learning curves of the synthesizability model trained on both the single-source and mixed-source datasets. It illustrates the model's accuracy on the training and validation sets over 500 epochs. Interestingly, the model exhibits different learning behaviors on the two datasets. Training on the mixed-source data results in rapid convergence, reaching a 90% accuracy on both the training and validation sets after the first epoch. The model exhibits minimal incremental learning over the subsequent epochs, and eventually reaches a near-perfect accuracy within less than 50 epochs. This rapid convergence suggests that the model distinguishes between synthesizable and unsynthesizable samples too easily, likely biased by the larger density of the unsynthesizable samples, rather than learning inherent synthesizability features. An apparent gap in prediction accuracy between the training and validation sets also suggests overfitting, although accuracy is surprisingly high on both sets. In contrast, when trained on the single-source data, the model exhibits a much slower convergence, starting with a modest 52% accuracy in the first epoch, which is as good as random guessing, before stabilizing at 80% after 100 epochs. Notably, the model trained on the single-source data does not show signs of over- or under-fitting, with both training and validation accuracy reaching 80%. During training, we apply a random rotation to each crystal image at each epoch to promote approximate rotation invariance (see details in [34]).

The disparity between the learning curves highlights that the classification success on the mixed-source data may be spurious and is likely influenced by inherent bias between class samples. In other words, we hypothesize that the bias in the training data has misled the model to learn features that are not truly indicative of the learning objective. The observations in this section elucidate that while the model's architecture is maintained, the nature and systematic differences within training examples can significantly influence model's learning behavior.

Evaluation on test set

We assess the performance of synthesizability models trained on both single-source and mixed-source data using their respective test sets. The synthesizability likelihood for each sample in the test set is averaged over an ensemble of 100 randomly rotated images of the crystal. Additional details on the choice of the ensemble size are provided in appendix A. For evaluating the classification performance on the test data, we employ metrics such as the area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and specificity (a classification threshold of 0.5 is utilized). Figures 3(c) and (d) compares the prediction performance of the model trained on mixed-source versus single-source data. In figure 3(c) the ROC curve for the mixed-source synthesizability model is illustrated, showing a high AUC value of 0.995. The mixed-source model exhibits high values for accuracy (0.967), precision (0.990), recall (0.971), and specificity (0.95), indicating the model's near-perfect performance in positive predictions as well as maintaining low false-negative predictions. The synthesizability likelihood distribution of test samples in the mixed-source model, as shown in figure 3(e), indicates that the model can distinctly differentiate between

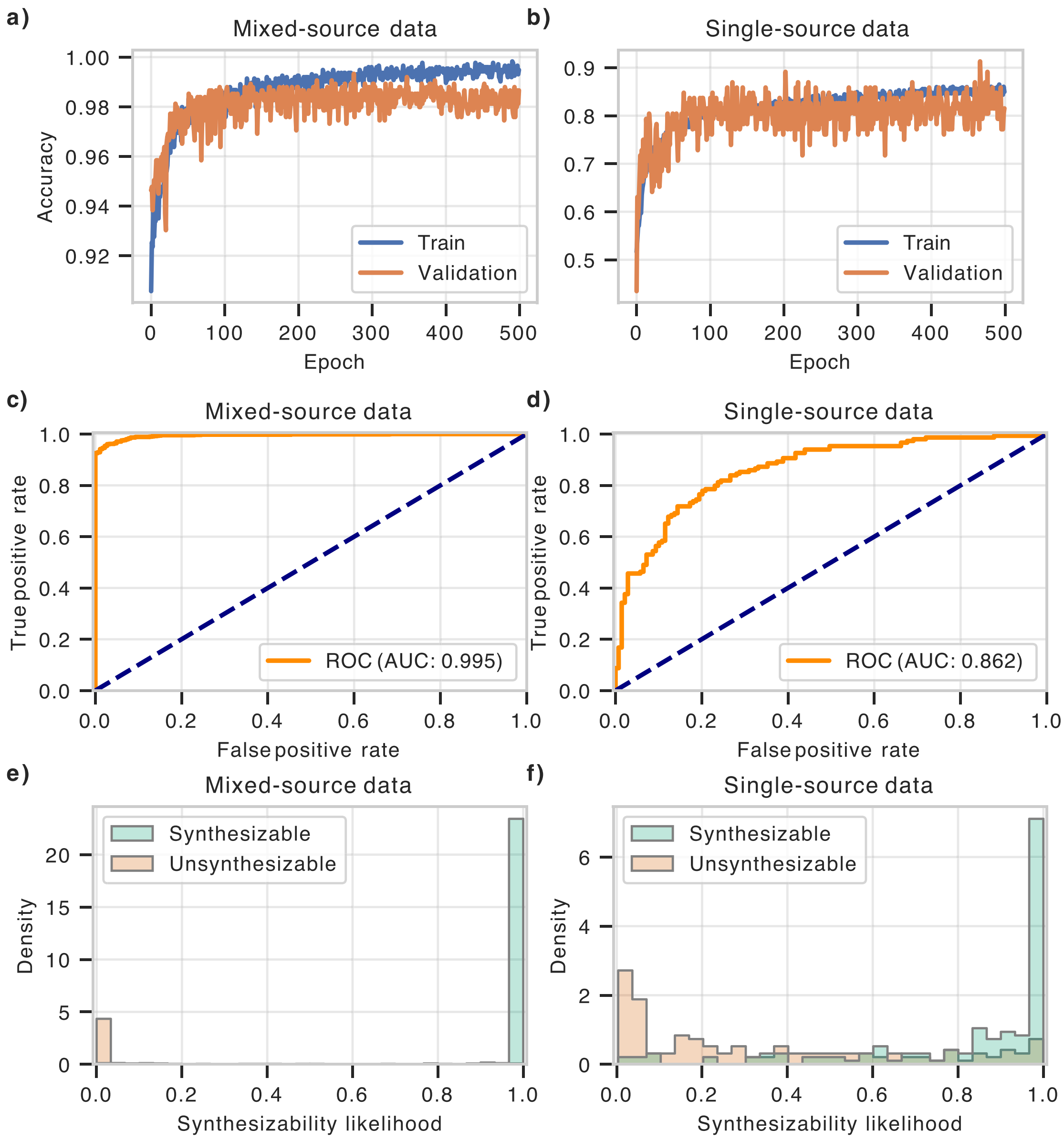

Figure 3. The synthesizability model's training history and performance using mixed-source and single-source data. (a), (b) Learning curves of the synthesizability model on the mixed-source and single-source data, respectively. The learning curves show the classification accuracy over epochs, for both training and validation sets. (c), (d) Receiver operating characteristic (ROC) curves of the synthesizability model on the mixed-source and single-source data, respectively. The area under the curve (AUC) is shown for each model. (e), (f) Normalized distribution of synthesizability likelihood of crystals in the test sets of the mixed-source and single-source data, respectively.

samples of the two classes. Almost all negative and positive samples are classified with synthesizability likelihoods of 0 and 1, respectively. This clear separation implies that the classification task is straightforward for the mixed-source model.

In figure 3(d), the ROC curve for the single-source model is presented. In comparison to the mixed-source model, the single-source model demonstrates an overall lower prediction performance, with an AUC of 0.862. The single-source model's accuracy is 0.778 (compared to mixed-source accuracy of 0.967). Recall remains reasonably high at 0.859 (compared to mixed-source accuracy of 0.971), indicating the single-source model's ability to detect positive samples with a low number of false negatives. The precision and specificity of 0.749 and 0.69 suggests a relatively higher number of false positive and negative predictions, respectively. Figure 3(f) illustrates the distribution of the single-source model's synthesizability likelihood, depicting the false positive and negative predictions.

The evaluation of the two models' performance on their respective test data reveals very similar trends to their performance on their training and validation sets. While both models are considered good classifiers

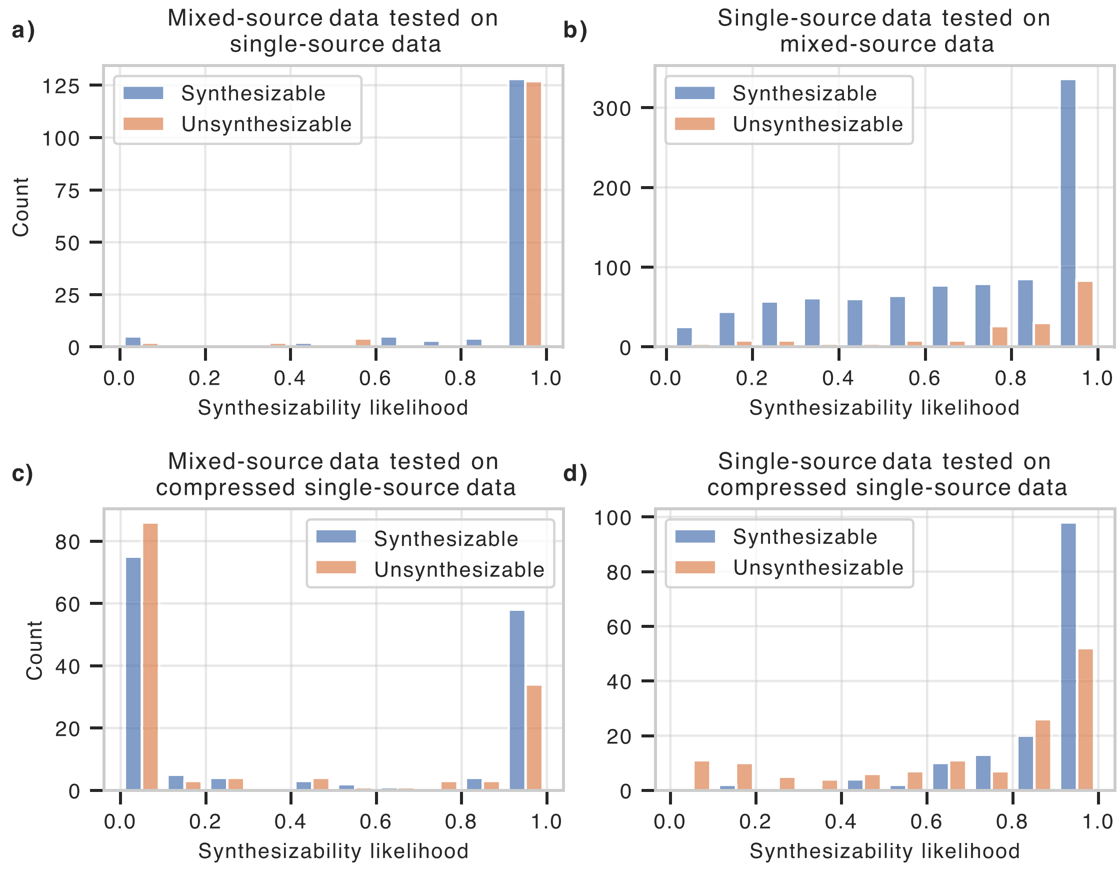

Figure 4. Evaluation of synthesizability model performance outside its test set under two scenarios: cross-evaluation on swapped test sets and assessment on compressed crystal data. Cross-evaluation assesses the performance of two models (mixed-source and single-source models) on each other's test sets. (a) Distribution of synthesizability likelihood for crystals in the single-source test set predicted by the mixed-source model, with accuracy, precision, recall, and specificity of 0.51, 0.51, 0.94, 0.04, respectively. (b) Distribution of synthesizability likelihood for crystals in the mixed-source test set predicted by the single-source model, with accuracy, precision, recall, and specificity of 0.63, 0.81, 0.73, 0.15, respectively. (c) Distribution of synthesizability likelihood for compressed crystals in the single-source test set predicted by the mixed-source model, with accuracy, precision, recall, and specificity of 0.56, 0.61, 0.43, 0.7, respectively. (d) Distribution of synthesizability likelihood for compressed crystals in the single-source test set predicted by the single-source model, with accuracy, precision, recall, and specificity of 0.61, 0.58, 0.94, 0.23, respectively.

according to key metrics, the mixed-source model demonstrates notably superior performance. In the next sections, we demonstrate that this superior performance is biased and thereby not reliable.

The evaluation exercise in this section demonstrates typical procedures widely used to assess machine learning model performances, specifically by evaluating the model's predictions on the test set. However, we illustrate that such evaluation procedures are not suitable for detecting spurious or biased learning by the model or the detrimental effects of data bias on the model's performance.

Cross-evaluation on swapped test sets

To extend beyond test set evaluation, we assess the classification metrics of both single-source and mixed-source models on swapped test sets. Specifically, we evaluate the mixed-source synthesizability model (trained on data from COD and CSPD) on the test set derived from single-source data (collected from MP). While the mixed-source synthesizability model performs near perfectly on its original test set, it encounters challenges in differentiating between unsynthesizable and synthesizable samples in the single-source test data. There is a significant performance drop in this scenario, with accuracy, precision, recall, and specificity dropping from 0.967 to 0.51, 0.990 to 0.51, 0.971 to 0.94, and 0.95 to 0.04, respectively. In figure 4(a), the distribution of synthesizability likelihood for single-source test samples is depicted using the mixed-source model. As observed in figure 4(a), the combination of a high number of false positives and a close-to-zero number of false negatives (high recall and very low specificity) implies that the model mistakenly identifies all crystal samples as synthesizable or positive. This observation can be explained by the fact that the mixed-source model tends to categorize any crystal sample within a density range of approximately 1.5–4 g/cc as synthesizable, a pattern learned from the bias in its training data (refer to figure 2(a)). This density range corresponds to the majority of samples in the single-source data (refer to figure 2(b)), including the test samples in this exercise. In essence, the mixed-source model has predominantly learned variations in crystal density rather than chemical and structural synthesizability attributes. This observation supports our hypothesis that the mixed-source model has captured a spurious correlation between density and synthesizability likelihood, derived from the bias in its training data. As demonstrated in the previous section, this spurious correlation cannot be detected through a standard test set evaluation procedure.

We also assess the performance of the single-source synthesizability model (trained on data exclusively from MP) on the test set derived from mixed-source data. As anticipated, the model's performance decreases compared to when evaluated on its original test set; however, the performance drop is less severe than that of the mixed-source model. The accuracy, precision, recall, and specificity of the single-source model are 0.63, 0.81, 0.73, 0.15, respectively, when tested on the mixed-source test data. Figure 4(b) depicts the distribution of synthesizability likelihood of mixed-source test samples using the single-source model. An interesting observation emerges: the recall is relatively high (0.73), indicating a low number of false negatives. In other words, the single-source model correctly predicts most of the synthesizable crystals as synthesizable. On the other hand, the specificity is low (0.15), indicating that many of the unsynthesizable samples are incorrectly predicted as synthesizable. The primary reason for the notable difference between recall and specificity is that the positive samples in the mixed-source data set share similarities in their density and mean atomic mass with both the positive and negative samples observed and learned by the single-source model (see figure 2). In simpler terms, the positive samples in the test set fall within the applicability domain of the single-source model. In contrast, the negative samples in the mixed-source dataset differ in their density and mean atomic mass from both the positive and negative samples learned by the single-source model, placing them outside the applicability domain of the model. This discrepancy explains the less accurate prediction of unsynthesizable samples. Kumagai *et al* demonstrated that the prediction performance of machine learning models significantly decreases outside their applicability domain [24].

Another important observation is that the prediction performance of the single-source model on mixed-source negative test samples is better than the mixed-source model's performance on single-source negative test samples. Specifically, the single source model's specificity on the mixed-source data is still larger than the mixed-source model's on the single source data (0.15 vs. 0.04), indicating that the single-source model generalizes better outside of its applicability domain. Although the number of mispredictions are high (attributed to test data being outside the applicability domain), the single-source model predicts a broader range of synthesizability likelihoods on negative samples compared to the mixed-source model tested on single-source data (compare figures 4(a) and (b)). This observation suggests that the single-source model does not recognize density as the primary correlated attribute to synthesizability. This further confirms our hypothesis that a model trained on heterogeneous or mixed-source data is susceptible to capturing spurious correlations.

Crystal compression test

To further test our hypothesis that the mixed-source model primarily bases its decisions on density, we conduct a compression experiment on crystal samples in the single-source test data (exclusively from MP). We then assess the performance of both the mixed-source and single-source models on the compressed crystal structures. For this experiment, we apply a 55% isotropic volume reduction to each crystal sample in the single-source test set. Figure 2(e) illustrates the density distribution of the test set samples before and after the volume reduction. As shown in figure 4(c), the mixed-source model predicts many of the samples as unsynthesizable. This result strongly supports our hypothesis because the same model could not recognize any of the same examples as unsynthesizable when uncompressed (compare figures 4(a) and (c)). In other words, the mixed-source model predictions largely vary with the density variation of the crystal sample, indicating that the model has established a strong correlation between crystal structure density and synthesizability. Additionally, on the compressed data, the mixed-source model mispredicts many of the positive samples as unsynthesizable when compressed (compare the blue bars in figures 4(a) and (c)). This showcases that the chemical or structural features underlying synthesizability of crystal compounds are not learned by the mixed-source model. On the other hand, the predictions of the single-source model on compressed single-source data did not change significantly compared to its predictions on the mixed-source data (compare figures 4(b) and (d)). This implies that the single-source model does not identify any strong correlation between crystal density and synthesizability likelihood. The performance of the single-source model drops on the compressed crystal test samples compared to the uncompressed test samples, with its accuracy reducing from 0.778 to 0.61 (compare figures 3(f) and 4(d)). However, this performance drop is much less significant compared to the mixed-source model, especially on the positive samples, showcasing the better generalization of the single-source model. As mentioned earlier, this performance drop is related to the use of the model outside its applicability domain. Even when utilized on the compressed crystal samples, the single-source model's performance on positive samples remains very good (recall of 0.94), indicating that the model has in fact learned the underlying chemical or structural features of synthesizability.

The compression test exercise in this section introduces a simple evaluation procedure for assessing how the performance of a model is influenced by the bias in data. In this study, we observed a clear bias in the density of crystal samples in the mixed-source data. Therefore, creating test data with systematic variation in their density through the compression test provides a reasonable platform to detect the propagated bias in the models' learning and performance.

4. Conclusion

This study compares the learning behavior and prediction performance of a machine learning model trained on two different data sets. The model performs a classification task to differentiate between samples of synthesizable versus unsynthesizable crystal compounds. To compile the data set for the binary classification, we follow two procedures: the mixed-source data set consists of data from separate computational and experimental crystal structure databases while the single-source data set consists of data from a single computational source. We detect a clear bias in the density and mean atomic mass distribution between samples of the two classes in the mixed-source data. Our results indicate that the mixed-source model produces biased and unreliable predictions, although all standard classification metrics suggest it is a near-perfect classifier. On the other hand, the single-source model shows a less accurate yet more reliable prediction performance. This study presents simple evaluation procedures beyond standard evaluation practices in the literature to measure the effect of data inconsistency on the model's prediction performance.

In conclusion, we underscore the potential for obscure inconsistency in data resulting from data collection across multiple sources. As demonstrated in this study, the collection of crystal structure samples for binary classification can introduce inconsistencies, particularly in their density distribution. This form of data bias is easily overlooked, given that it remains undetectable through standard evaluation procedures commonly employed in typical machine learning studies for materials prediction or discovery. Hence, it becomes imperative to systematically compare data across various properties, such as density, chemical composition, and structural distribution, in order to identify any potential imbalances or sources of bias before initiating model training.

While demonstrating the detrimental effect of data bias on machine learning models, this study does not propose detailed pathways for mitigating such bias in the model's learning and performance. However, we do highlight the superior performance and generalization of a model trained on homogeneous data. Mitigation strategies should be tailored to the specific nature of the data bias and the chosen learning approach. In the case of this study, using universal potential models, like CHGNET [59], can provide computationally feasible volume-relaxation tool to address the density bias in CSPD-derived structures (see appendix F). However, caution is advised when using these models, as they are not specifically designed to describe forces in highly distorted structures far from equilibrium, which is often the case for CSPD-derived structures or those generated by similar high-throughput structure generation methods. If computationally feasible, more accurate models, such as DFT, are recommended. More broadly, the path forward should involve careful data selection methods that promote diverse and unbiased datasets. Mitigating data bias will be crucial for achieving reliable and useful machine learning predictions in materials science. Data bias detection and mitigation studies in the literature [24–28], while few, provide the roadmap for future studies in machine learning applied to materials science.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/kadkhodaei-research-group/XIE-SPP.

Acknowledgment

This research is based upon work supported by the National Science Foundation (NSF) under Award Numbers DMR-2119308 and DMR-2119065. We used resources at the Electronic Visualization Laboratory (EVL) at University of Illinois Chicago, available through the NSF Award CNS-1828265.

Code availability

The codes developed or utilized in this study are openly accessible to support transparency and facilitate further research. They can be found in our GitHub repository at https://github.com/kadkhodaei-research-group/XIE-SPP.

Author Contribution Statement

A D and S K conceived the research. S K and E Z supervised the study. A D carried out the method development and performed the calculations and analysis. All authors participated in discussing the results and commented on the manuscript.

Competing Interests

All authors declare no financial or non-financial competing interests.

Appendix A. Rotational ensemble averaging

Table A.1. The agreement between predicted labels of the 288 samples of the single-source model test set in terms of the ensemble size used for each sample prediction. A sample is deemed consistent if all predicted labels are identical, and inconsistent otherwise. The consistent prediction percentage column represents the fraction of samples with consistent predicted labels among the 288 samples. The standard deviation of an inconsistent sample is measured among all the copies in the ensemble. The average standard deviation over inconsistent samples is presented in the last column.

Ensemble size	Number of inconsistent predictions	Consistent prediction percentage	Average standard deviation of inconsistent predictions
1	123	56.4%	0.109
2	79	72.0%	0.085
5	45	84.0%	0.055
10	39	86.2%	0.039
20	33	88.3%	0.030
30	27	90.4%	0.024
50	18	93.6%	0.020
75	17	94.0%	0.017
100	12	95.7%	0.015
150	11	96.1%	0.011
200	9	96.8%	0.010

We employ an ensemble averaging method for predicting the synthesizability likelihood of crystal compounds. Once a crystal sample is input into the trained model, a ensemble of N randomly rotated instances of the sample is generated and the synthesizability likelihood prediction is averaged over the ensemble, where N is an adjustable parameter of the model denoting the ensemble size. The ensemble averaging methods improves the prediction accuracy and robustness of our model, as shown in our earlier study for formation energy prediction [34], and as illustrated in the following for synthesizability prediction.

Table A.1 illustrates the single-source model's prediction consistency and stability on its test set in terms of the ensemble size. As the ensemble size increases, the number of inconsistent predictions dramatically decreases. The prediction on a crystal sample is considered consistent if the predicted labels for all the rotational instances of the crystal sample are identical, and inconsistent otherwise. A negative label is assigned if the predicted synthesizability likelihood is below 0.5 and a positive label is assigned otherwise (i.e. a classification threshold of 0.5 is utilized). There exist 288 distinct crystal samples in the single-source model test set. As shown in table A.1, for an ensemble of size 1, 123 out of the 288 samples have inconsistent predictions, resulting in consistent prediction percentage of around 56%. A rapid monotonic increase in prediction consistency is observed as the ensemble size increases, leading to a consistency percentage of 90% and above 96% for ensemble sizes of 30 and 100, respectively. Another metric we use to measure the consistency is the standard deviation of synthesizability likelihood predictions for a given crystal sample. The standard deviation is measured over N rotational instances of the crystal compound. Table A.1 shows the average standard deviation of synthesizability likelihood of the inconsistent crystal samples in the test set in terms of the ensemble size. A descending trend in the standard deviation as the ensemble size grows is demonstrative of the model's enhanced confidence in its predictions. This suggests that larger ensembles translate to more consistent and confident results. For example, for an ensemble of size 50, the average standard deviation over the 11 inconsistent crystal samples stands at a mere 0.011. Notably, the small standard deviation indicates that the inconsistent predictions corresponds to crystal samples located close to the decision boundary, where the predictions should be close to 0.5 with small oscillations of around 0.011 that render their labels positive or negative.

We select an ensemble size of 100 for our model prediction. For an ensemble size of 100, the model exhibits a high consistency percentage of 95.7% on its test samples, while simultaneously maintaining a low average standard deviation of 0.015 over the inconsistent samples. Beyond ensemble 100, the incremental improvements in both the consistency percentage and average standard deviation are relatively minor. Given these results, ensemble 100 emerges as an optimal balance between performance and computational efficiency. It offers high consistency and confidence in predictions without the necessity for considerably larger ensemble sizes that would demand more computational resources without a sufficient improvement in model performance.

Appendix B. Well-studied chemical compositions in the materials science literature

We utilize a natural language processing model developed by Tshitoyan *et al* [60] that encompasses knowledge from the materials science literature spanning from 1922 to 2018. We rank the available chemical compositions in the literature based on their frequency of occurrence. The top 0.1% comprises the first 108 unique compositions, each of which is mentioned at least 3306 times (see supplementary table 1 in [18]). These compositions are shown in table B.1.

Table B.1. The top 0.1% chemical compositions in the materials science literature based on their frequency of occurrence, collected using the natural language processing model of [60]. Formulas are ordered by frequency, decreasing from top left to bottom right, with frequency decreasing left to right within each row.

O_2Ti	OZn	CO_2	O_2Si	Al_2O_3	CSi
$AsGa$	GaN	O_2Zr	O_2Sn	MgO	CdS
CeO_2	H_2O	Fe_3O_4	$ClNa$	CuO	NiO
CH_4	SZn	NTi	O_3W	MoS_2	AlN
H_3N	$CdTe$	Fe_2O_3	ClH	CTi	O_3Y_2
MnO_2	$HNaO$	CaO	InP	Co_3O_4	BaO_3Ti
$CdSe$	N_4Si_3	Cu_2O	$AsIn$	O_5V_2	O_2S
O_3SrTi	H_2O_4S	$SeZn$	NO_2	H_2S	MoO_3
$NiTi$	$FeLiO_4P$	ClK	B_2Ti	$AsGaIn$	$GeSi$
Cr_2O_3	In_2O_3	B_2O_3	$AgCl$	HfO_2	CsN
$AlAsGa$	N_2O	Bi_2O_3	B_2Mg	GeV	La_2O_3
HNO_3	FLi	$AlNi$	$CCaO_3$	CaF_2	Nb_2O_5
O_2U	$GaInN$	PbS	$AlGaN$	CH_2	B_4C
$AlTi$	O_4S	$BiFeO_3$	$LiMn_2O_4$	Na_2O	$CoLiO_2$
$CoFe_2O_4$	O_2Ru	$LiNbO_3$	$CsFOS$	B_2Zr	$GaSb$
OPb	BRh	CoO	CrN	Na_2O_4S	MnO
CH_3	BiO_4V	O_2V	H_2Mg	$Li_4O_12Ti_5$	O_5Ta_2
O_5P_2	$FePt$	Li_2O	FeO	NO_3	$TeZn$

Appendix C. CSPD atomic structure generator

The CSPD[52] is constructed by extracting crystal structure prototypes from the COD[61], which includes various compounds except bio-polymers. The process involves filtering structures for quality, classifying them based on composition and atom count, distinguishing between inorganic and organic structures, and assessing structural similarity.

The CSPD approach, developed by Chuanxun Su *et al*, generates structures for prediction by first transforming a raw database into a composition-crystal-structure prototype database, significantly reducing the number of candidate structures [52]. This process, known as the big data method (BDM), involves selecting structure prototypes based on targeted composition, substituting elements, adjusting lattice parameters, and checking minimal interatomic distances. The effectiveness of this method was demonstrated through the generation and evaluation of candidate structures for typical systems, showing that the BDM not only efficiently predicts lower-energy structures but also aligns with experimental findings, underscoring its potential in identifying novel, energetically favorable materials configurations.

The CSPD structure generator is publicly available on GitHub [62]. When utilizing the CSPD for structure generation, users can choose to specify either atomic density or volume as guiding parameters for crystal structure creation. In instances where users do not provide these parameters, CSPD employs the covalent radii of species to deduce atomic density, determining the arrangement density of atoms within the generated structure. This approach provides flexibility in tailoring structure generation without requiring explicit pressure settings. In this study, we have used this option to generate crystal samples for the unsynthesizable class.

Alternative methods to the CSPD have been developed in recent years. One notable examples is the CrystaLLM by Luis Antunes *et al*, which presents an innovative approach to the generation of crystal structures [63]. Utilizing autoregressive large language models to interpret Crystallographic Information File formats, CrystaLLM accelerates the discovery of inorganic compounds suitable for applications in energy and electronics, highlighting a significant advancement in the efficiency of crystal structure prediction.

Appendix D. Distribution of atomic species across datasets

Figure D.1 illustrates the the distribution of atomic species between the synthesizable and unsynthesizable classes in both the mixed-source and single-source data.

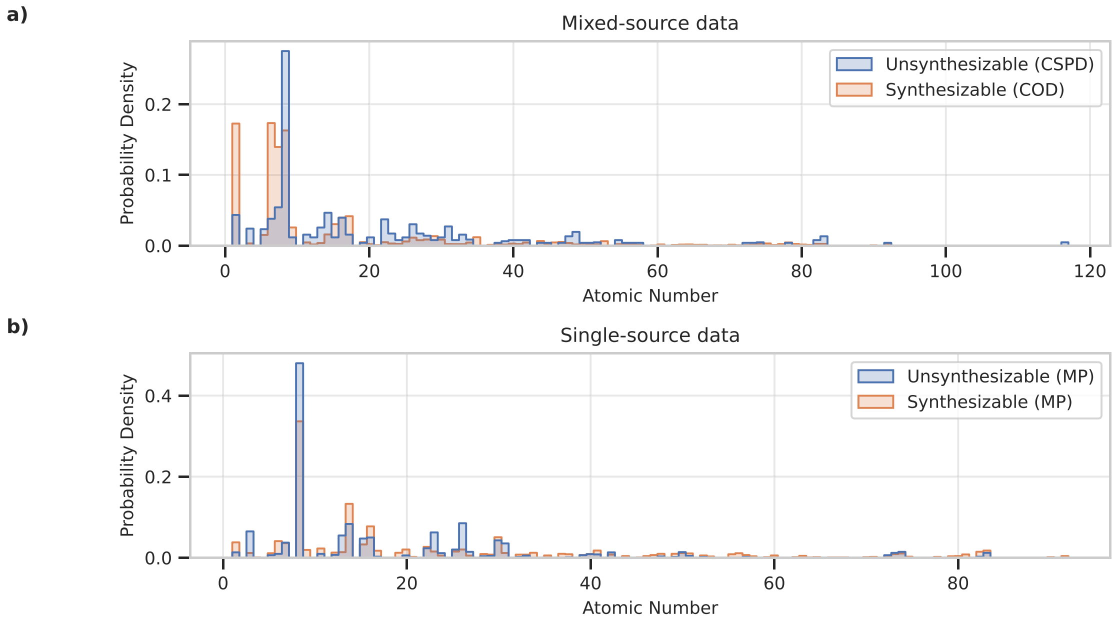

Figure D.1. Distribution for atomic species among the synthesizable and unsynthesizable samples in the mixed-source and single-source databases.

Appendix E. Distribution of ICSD crystals over external pressure

See figure E.1.

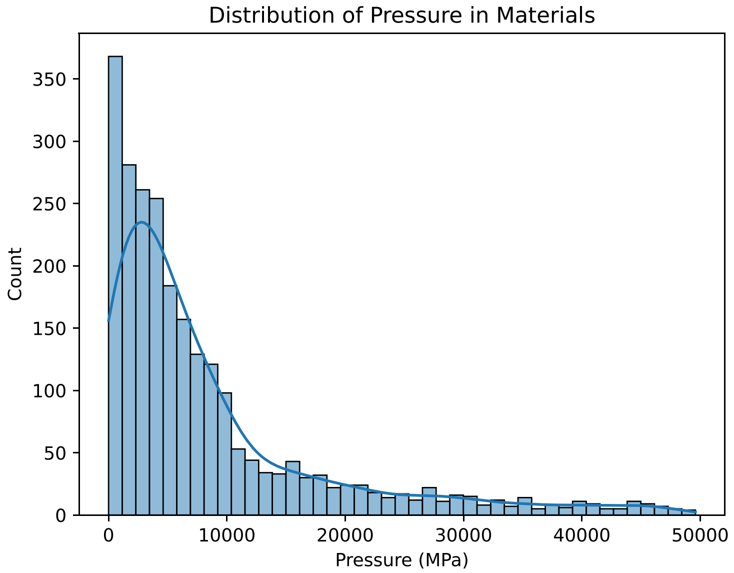

Figure E.1. Distribution for high-pressure ICSD crystal samples (i.e. above ambient pressure). Only samples with pressure below 50 GPa are displayed. The pressure value indicates the external pressure during synthesis.

Appendix F. Mitigating density bias in the mixed-source data

We employ CHGNET [59] to relax the structures in the mixed-source dataset. The relaxation process is conducted using the FIRE optimizer, which adjusts the crystal cell and atomic positions until the CHGNET-calculated forces converge to 0.1 eV Å^{-1}, or until the optimizer reaches 300 iterations. Negative samples from the CSPD are uniformly expanded by 250% before being subjected to optimization. This manual expansion helps address the inherent bias of CSPD-derived structures toward small atomic volumes, making them more suitable for optimization with CHGNET and resulting in a larger proportion of negative samples being successfully optimized. Despite these efforts, CHGNET failed to optimize a significant portion of the dataset, with 295 out of 600 CSPD samples and 1621 out of 3000 COD samples not converging (i.e. the forces diverged even after 300 iterations). Figure F.1 shows the distribution of the maximum atomic forces at the final step for the successfully relaxed structures. Additionally, figure F.2 presents the density and atomic volume distribution of the mixed-source data before and after CHGNET-based relaxation. Following the relaxation process, the disparity in density and atomic volume distributions between synthesizable and non-synthesizable samples was reduced.

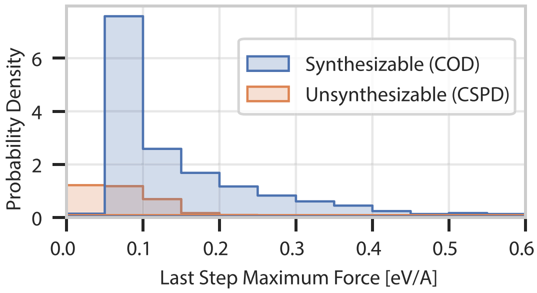

Figure F.1. Distribution of the last relaxation step atomic force for the synthesizable and unsynthesizable crystal samples in the mixed source dataset.

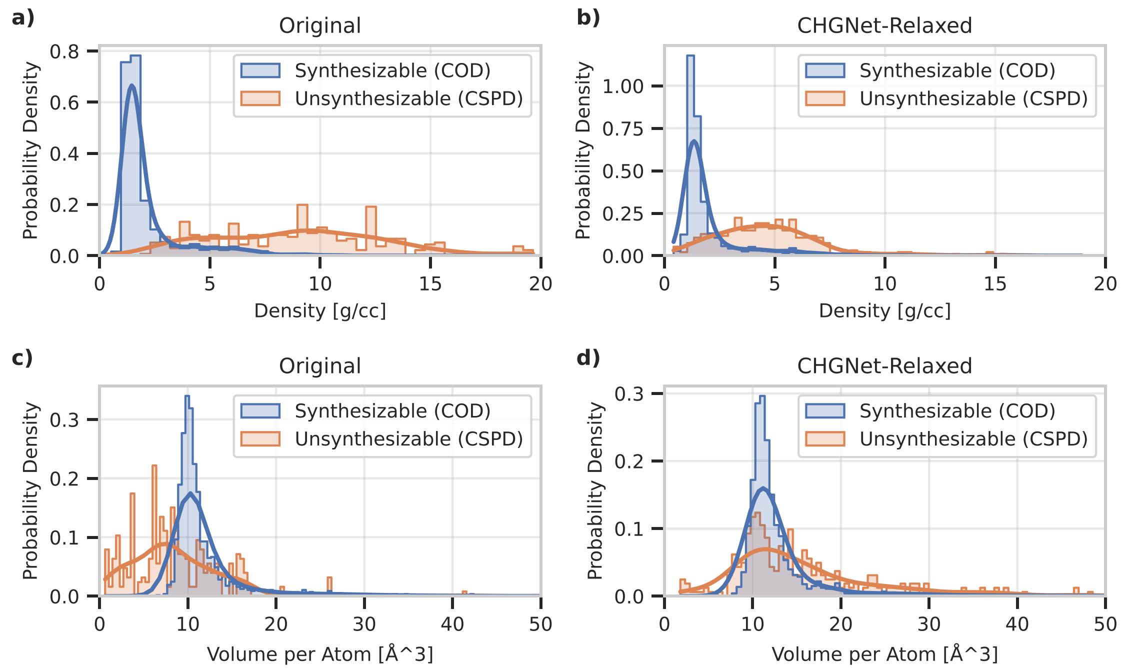

Figure F.2. Density and atomic volume distributions of crystal samples in the mixed source dataset before and after CHGNET relaxation.

ORCID iDs

Samad Hajinazar https://orcid.org/0000-0002-7255-5932
Eva Zurek https://orcid.org/0000-0003-0738-867X
Sara Kadkhodaei https://orcid.org/0000-0003-4263-3191

References

[1] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55

[2] Himanen L, Geurts A, Foster A S and Rinke P 2019 Data-driven materials science *Adv. Sci.* **1900808** 23

[3] de Pablo J J *et al* 2019 New frontiers for the materials genome initiative *npj Comput. Mater.* **5** 41

[4] Tian Y, Xue D, Yuan R, Zhou Y, Ding X, Sun J and Lookman T 2021 Efficient estimation of material property curves and surfaces via active learning *Phys. Rev. Mater.* **5** 013802

[5] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S and Tropsha A 2017 Universal fragment descriptors for predicting properties of inorganic crystals *Nat. Commun.* **8** 15679

[6] Gossett E *et al* 2018 Aflow-ml: a restful api for machine-learning predictions of materials properties *Comput. Mater. Sci.* **152** 134–45

[7] Umehara M, Stein H S, Guevarra D, Newhouse P F, Boyd D A and Gregoire J M 2019 Analyzing machine learning models to accelerate generation of fundamental materials insights *npj Comput. Mater.* **5** 34

[8] Jablonka K M, Ongari D, Moosavi S M and Smit B 2020 Big-data science in porous materials: materials genomics and machine learning *Chem. Rev.* **120** 8066–129

[9] Himanen L, Jäger M O, Morooka E V, Federici Canova F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 Dscribe: library of descriptors for machine learning in materials science *Comput. Phys. Commun.* **247** 106949

[10] Morgan D and Jacobs R 2020 Opportunities and challenges for machine learning in materials science *Annu. Rev. Mater. Res.* **50** 71–103

[11] Hart G L W, Mueller T, Toher C and Curtarolo S 2021 Machine learning for alloys *Nat. Rev. Mater.* **6** 730–55

[12] Gong S, Wang S, Zhu T, Chen X, Yang Z, Buehler M J, Shao-Horn Y and Grossman J C 2021 Screening and understanding Li adsorption on two-dimensional metallic materials by learning physics and physics-simplified learning *JACS Au* **1** 1904–14

[13] Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A and Han T Y-J 2022 Explainable machine learning in materials science *npj Comput. Mater.* **8** 204

[14] Damewood J, Karaguesian J, Lunger J R, Tan A R, Xie M, Peng J and Gómez-Bombarelli R 2023 Representations of materials for machine learning *Annu. Rev. Mater. Res.* **53** 399–426

[15] Xu P, Ji X, Li M and Lu W 2023 Small data machine learning in materials science *npj Comput. Mater.* **9** 42

[16] Agrawal A and Choudhary A 2016 Perspective: materials informatics and big data: realization of the 'fourth paradigm' of science in materials science *APL Mater.* **4** 053208

[17] Liang W, Tadesse G A, Ho D, Fei-Fei L, Zaharia M, Zhang C and Zou J 2022 Advances, challenges and opportunities in creating data for trustworthy AI *Nat. Mach. Intell.* **4** 669–77

[18] Davariashtiyani A, Kadkhodaie Z and Kadkhodaei S 2021 Predicting synthesizability of crystalline materials via deep learning *Commun. Mater.* **2** 115

[19] Frey N C, Wang J, Vega Bellido G I, Anasori B, Gogotsi Y and Shenoy V B 2019 Prediction of synthesis of 2D metal carbides and nitrides (mxenes) and their precursors with positive and unlabeled machine learning *ACS Nano* **13** 3031–41

[20] Jang J, Gu G H, Noh J, Kim J and Jung Y 2020 Structure-based synthesizability prediction of crystals using partially supervised learning *J. Am. Chem. Soc.* **142** 18836–43

[21] Antoniuk E R, Cheon G, Wang G, Bernstein D, Cai W and Reed E J 2023 Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions *npj Comput. Mater.* **9** 155

[22] Gleaves D, Fu N, Dilanga Siriwardane E M, Zhao Y and Hu J 2023 Materials synthesizability and stability prediction using a semi-supervised teacher-student dual neural network *Digit. Discov.* **2** 377–91

[23] Gong C, Wang Q, Liu T, Han B, You J, Yang J and Tao D 2022 Instance-dependent positive and unlabeled learning with labeling bias estimation *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 4163–77

[24] Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y and Kurosaki K 2022 Effects of data bias on machine-learning–based material discovery using experimental property data *Sci. Technol. Adv. Mater.: Methods* **2** 302–9

[25] Zhang H, Chen W W, Rondinelli J M and Chen W 2023 ET-AL: entropy-targeted active learning for bias mitigation in materials data *Appl. Phys. Rev.* **10** 021403

[26] Li K, DeCost B, Choudhary K, Greenwood M and Hattrick-Simpers J 2023 A critical examination of robustness and generalizability of machine learning prediction of materials properties *npj Comput. Mater.* **9** 55

[27] Zhang Y and Ling C 2018 A strategy to apply machine learning to small datasets in materials science *npj Comput. Mater.* **4** 25

[28] Breuck P-P D, Evans M L and Rignanese G-M 2021 Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODnet *J. Phys.: Condens. Matter.* **33** 404002

[29] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501–9

[30] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies *npj Comput. Mater.* **1** 15010

[31] Choudhary K *et al* 2020 The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design *npj Comput. Mater.* **6** 173

[32] Choudhary K and DeCost B 2021 Atomistic line graph neural network for improved materials property predictions *npj Comput. Mater.* **7** 185

[33] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002

[34] Davariashtiyani A and Kadkhodaei S 2023 Formation energy prediction of crystalline compounds using deep convolutional network learning on voxel image representation *Commun. Mater.* **4** 105

[35] Jiang Y, Chen D, Chen X, Li T, Wei G-W and Pan F 2021 Topological representations of crystalline compounds for the machine-learning prediction of materials properties *npj Comput. Mater.* **7** 28

[36] Jones E B and Stevanović V 2017 Polymorphism in elemental silicon: probabilistic interpretation of the realizability of metastable structures *Phys. Rev.* B **96** 184101

[37] Zhu R, Tian S I P, Ren Z, Li J, Buonassisi T and Hippalgaonkar K 2023 Predicting synthesizability using machine learning on databases of existing inorganic materials *ACS Omega* **8** 8210–8

[38] Raccuglia P, Elbert K C, Adler P D F, Falk C, Wenny M B, Mollo A, Zeller M, Friedler S A, Schrier J and Norquist A J 2016 Machine-learning-assisted materials discovery using failed experiments *Nature* **533** 73–76

[39] Kim E, Huang K, Jegelka S and Olivetti E 2017 Virtual screening of inorganic materials synthesis parameters with deep learning *npj Comput. Mater.* **3** 53

[40] Huo H *et al* 2019 Semi-supervised machine-learning classification of materials synthesis procedures *npj Comput. Mater.* **5** 62

[41] Kononova O, Huo H, He T, Rong Z, Botari T, Sun W, Tshitoyan V and Ceder G 2019 Text-mined dataset of inorganic materials synthesis recipes *Sci. Data* **6** 203

[42] Kim E *et al* 2020 Inorganic materials synthesis planning with literature-trained neural networks *J. Chem. Inf. Model* **60** 1194–201

[43] Karpovich C, Jensen Z, Venugopal V and Olivetti E, 2021 Inorganic synthesis reaction condition prediction with generative machine learning (arXiv:2112.09612)

[44] Wang Z *et al* 2022 Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature *Sci. Data* **9** 231

[45] Huo H, Bartel C, He T, Trewartha A, Dunn A, Ouyang B, Jain A and Ceder G 2022 Machine-learning rationalization and prediction of solid-state synthesis conditions *Chem. Mater.* **34** 7323–36

[46] Karpovich C, Pan E, Jensen Z and Olivetti E 2023 Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction *Chem. Mater.* **35** 1062–79

[47] McDermott M J *et al* 2023 Assessing thermodynamic selectivity of solid-state reactions for the predictive synthesis of inorganic materials *ACS Cent. Sci.* **9** 1957–75

[48] Aykol M, Hegde V I, Hung L, Suram S, Herring P, Wolverton C and Hummelshøj J S 2019 Network analysis of synthesizable materials discovery *Nat. Commun.* **10** 2018

[49] Aykol M, Montoya J H and Hummelshøj J 2021 Rational solid-state synthesis routes for inorganic materials *J. Am. Chem. Soc.* **143** 9244–59

[50] McDermott M J, Dwaraknath S S and Persson K A 2021 A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis *Nat. Commun.* **12** 3097

[51] Gražulis S, Chateigner D, Downs R T, Yokochi A F T, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P and Le Bail A 2009 Crystallography Open Database – an open-access collection of crystal structures *J. Appl. Crystallogr.* **42** 726–9

[52] Su C, Lv J, Li Q, Wang H, Zhang L, Wang Y and Ma Y 2017 Construction of crystal structure prototype database: methods and applications *J. Phys.: Condens. Matter* **29** 165901

[53] Zagorac D, Müller H, Ruehl S, Zagorac J and Rehme S 2019 Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features *J. Appl. Crystallogr.* **52** 918–25

[54] Quirós M, Gražulis S, Girdzijauskaitė S, Merkys A and Vaitkus A 2018 Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database *J. Cheminf.* **10** 23

[55] Merkys A, Vaitkus A, Butkus J, Okulič-Kazarinas M, Kairys V and Gražulis S 2016 COD::CIF::Parser: an error-correcting CIF parser for the Perl language *J. Appl. Crystallogr.* **49** 292-301

[56] Gražulis S, Merkys A, Vaitkus A and Okulič-Kazarinas M 2015 Computing stoichiometric molecular composition from crystal structures *J. Appl. Crystallogr.* **48** 85–91

[57] Gražulis S, Davskevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, Serebryanaya N R, Moeck P, Downs R T and Le Bail A 2012 Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration *Nucleic Acids Res.* **40** D420–7

[58] Downs R T and Hall-Wallace M 2003 The American mineralogist crystal structure database *Am. Mineral.* **88** 247–50

[59] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel C J and Ceder G 2023 Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling *Nat. Mach Intell.* **5** 1–11

[60] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–98

[61] Vaitkus A, Merkys A, Sander T, Quirós M, Thiessen P A, Bolton E E and Gražulis S 2023 A workflow for deriving chemical entities from crystallographic data and its application to the crystallography open database *J. Cheminf.* **15** 123

[62] Su C, 2018 Atomic structure generator (available at: https://github.com/SUNCAT-Center/AtomicStructureGenerator.git)

[63] Antunes L M, Butler K T and Grau-Crespo R, 2024 Crystal structure generation with autoregressive large language modeling (arXiv:2307.04340)