ScGAN: a generative adversarial network to predict hypothetical superconductors

Evan Kim and S V Dordevic

1 Tesla STEM High School, Redmond, WA 98053, United States of America
2 Department of Physics, The University of Akron, Akron, OH 44325, United States of America

E-mail: dsasa@uakron.edu

Received 25 June 2023, revised 20 September 2023
Accepted for publication 27 September 2023
Published 11 October 2023

Abstract
Despite having been discovered more than three decades ago, high temperature superconductors (HTSs) lack both an explanation for their mechanisms and a systematic way to search for them. To aid this search, this project proposes ScGAN, a generative adversarial network (GAN) to efficiently predict new superconductors. ScGAN was trained on compounds in Open Quantum Materials Database and then transfer learned onto the SuperCon database or a subset of it. Once trained, the GAN was used to predict superconducting candidates, and approximately 70% of them were determined to be superconducting by a classification model—a 23-fold increase in discovery rate compared to manual search methods. Furthermore, more than 99% of predictions were novel materials, demonstrating that ScGAN was able to potentially predict completely new superconductors, including several promising HTS candidates. This project presents a novel, efficient way to search for new superconductors, which may be used in technological applications or provide insight into the unsolved problem of high temperature superconductivity.

Keywords: ScGAN, GAN, machine learning, superconductors

1. Introduction
In recent years, superconductors have been applied in a variety of important technologies such as power transmission lines, MRI magnets, Maglev trains, and quantum computers. Quantum computers are especially important as they are expected to solve problems that are too computationally expensive for current classical computers. However, to be used in these applications the superconductors must be cooled below their critical temperatures ($T_c$), which for most current superconductors are very low. For instance, Google’s quantum computer must be maintained at 0.02 K, which severely limits its general use [1]. This points to a growing need for superconductors with higher $T_c$, which has made them an active research topic for the last couple of decades.

The mechanisms of superconductivity in most materials with relatively high $T_c$ is not fully understood, which means there is no systematic way to search for new materials or to predict their critical temperatures [2]. Thus, the current procedure for finding high temperature superconductors (HTSs) is essentially trial–and–error, which is extremely inefficient. This was exemplified in a recent study, which found that only about 3% of the approximately 1000 materials surveyed were superconducting [3]. Furthermore, the study failed to find any superconductors with $T_c > 60$ K. This extreme inefficiency means that the likelihood of manually finding new superconductors, especially HTSs, is extremely low.

To address this difficulty, the use of computational tools in superconductivity research has become popular in recent
years [4]. In particular, there have been several studies utilizing machine learning to predict whether a given chemical compound will be superconducting or not. The earliest such study was [5] in which two random forest-based models were built: a classification model for predicting superconductivity and a regression model for predicting superconducting transition temperature [5]. The models were successfully applied on the SuperCon database, achieving a 92% accuracy in classification and a $R^2 = 0.88$ for regression. They did run across one limitation of their machine learning model, however. When trained on a certain class of superconductors (e.g. cuprates), the model was unable to make good predictions on other classes of superconductors (e.g. pnictides).

Following this pioneering study, there have been several other studies applying machine learning methods, such as a $K$-nearest neighbors (KNNs) algorithm [6] and a deep learning model [7]. The KNNs algorithm reported improvements on the previous study from [5] in terms of overall performance: an $R^2$ of 0.93 and a classification accuracy of 96.5%. The deep learning model, on the other hand, showed that it might be possible to overcome the limitation faced in [5], as they were able to make predictions about pnictide superconductors from the training data that did not contain them.

The way each of these previous studies predicted new superconducting materials was by running their trained model on a database of known existing chemical compounds and finding which compounds the model indicated could be superconducting. This procedure has several notable limitations. First, these studies miss out on the vast chemical composition space that is not already contained in existing databases. Second, commonly used databases (such as ICSD and Open Quantum Materials Database (OQMD)) contain mostly stoichiometric compounds, whereas many superconductors are non-stoichiometric (cuprates and pnictides, in particular), and so many possibilities were missed in that manner as well (see for example table 3 in [5] and table 1 in [6]). Finally, the discovery of superconductors in the lab usually does not happen that way; they are discovered by synthesizing new materials and testing them, rather than checking the known ones.

In order to overcome these limitations, in this work we employ generative adversarial networks (GANs) [8]. Generative models refer to a general class of machine learning models which are able to generate things that resemble the input dataset. They have proven to be extremely powerful, and in recent years have found numerous applications such as science, engineering, medicine, art, video games, deepfakes, etc [9]. In most cases GANs performed better than other generative models, such as variational autoencoders, because they are able to learn more hidden rules in the input data set. For example, in [10] a GAN model which generated new chemical compounds with 92.53% novelty and 84.5% validity was reported. Another work, [11] applied GANs to the SuperCon dataset, but for the purpose of characterizing rather than prediction.

In this work, we combine the general idea of GANs with the previous superconductor models, and propose the first GAN to predict new superconductors (ScGAN). Our ScGANs are based on chemical composition only, and are able to generate new superconducting materials with high novelty and validity. The reason we choose to use chemical composition only is that it has been shown that additional descriptors do not lead to additional improvements [6] and many descriptors are unavailable in the database. The paper is organized as follows. In section 2 we present the details of the creation of the ScGANs. In section 3 the main results of the study are discussed. In particular, we present a list of hypothetical superconducting materials generated by our ScGANs, as well as their predicted critical temperatures. Finally, in section 4 we summarize the most important findings.

2. Methodology

As stated in the introduction, we chose the GAN as our generative model. Its structure is shown in figure 1. A GAN is composed of two competing neural networks, the generator and discriminator. The generator takes in random noise and generates a ‘fake’ compound, which the discriminator attempts to determine if it is real or not using existing data of real compounds. Each of them then updates their parameters based on the performance of the discriminator. The two networks improve their performance iteratively until a generator can generate realistic looking compounds, while the discriminator can detect unrealistic looking ones [8].

Figure 2 depicts the training/testing process of our GANs and it has three main stages. The main idea is to have the model first learn general chemical composition rules by training off of a larger dataset (OQMD database) of general chemical compounds and then transfer learning it onto a (much smaller) dataset of known superconducting materials (SuperCon). The idea of transfer learning is to allow a model to learn even with a limited amount of data, which is the case here as the general compounds dataset (OQMD) has on the order of $\sim 10^6$ data points, while the superconductor datasets will only have on the order of $\sim 10^4$ data points.

2.1. Data collection

Data were sourced from two open-source databases: SuperCon [12] and the OQMD [13, 14]. SuperCon is the largest database for superconducting materials with around 30000 superconductors before filtering. Similar to what was done in some previous studies [6, 15], we only used the chemical compositions of the materials extracted from SuperCon. On the other hand, OQMD is a much larger database with around $10^6$ DFT-calculated compounds, most of which are not superconductors.

2.2. Data processing

Before filtering the data, all compounds had to be expressed in a common format so that unwanted datapoints could be detected regardless of small differences in formatting across
Figure 1. The architecture of the GAN model used. $Y_{1.2}Ba_{0.8}CuO_4$ and $O_{7.25}Ca_{0.62}Cu_{2.13}Sr_{1.91}Y_{0.37}Bi_{1.71}$ are examples of real (from SuperCon) and ‘fake’ superconductors, respectively.

Figure 2. The overall training/testing process from the data to the final GAN model. It is composed of three main stages: (a) data processing, (b) training on the OQMD dataset, and finally (c) transfer learning onto the superconductor dataset (the whole SuperCon, or a part of it).

Figure 3. Chemical composition data represented as a matrix [6] in $\mathbb{R}^{96 \times N}$, where $N$ is the number of datapoints. Each column is a single compound, with each entry representing the number of each element present in the compound. Note that the numbers in the matrix are for illustration purposes only; they do not represent any real compounds or superconductors.

Table 1. Distribution of the superconductors in our filtered version of SuperCon by class. Each of these served as different training sets. We notice that the number of pnictide entries is much smaller compared with cuprates and others.

| Class     | Quantity | Percentage |
|-----------|----------|------------|
| Cuprates  | 7304     | 44.4%      |
| Pnictides | 1436     | 8.7%       |
| Others    | 7749     | 47.0%      |
| Everything| 16489    | 100%       |
A GAN is a type of generative model that has two competing neural networks, a generator and a discriminator, as shown in figure 1. Traditional GANs, however, can suffer from issues such as mode collapse and gradient vanishing, so the Wasserstein GAN with gradient penalty [16] was used instead. In the Wasserstein GAN with gradient penalty, the loss functions are

\[
\text{Loss}_D = \mathbb{E}_{\tilde{x}\sim P_d} [D(\tilde{x})] - \mathbb{E}_{x\sim P_g} [D(x)] + \lambda \mathbb{E}_{\tilde{x}\sim P_d} \left[ (\|\nabla_{\tilde{x}} D(\tilde{x})\| - 1)^2 \right],
\]

\[
\text{Loss}_G = -\mathbb{E}_{\tilde{x}\sim P_d} [D(\tilde{x})].
\]

Here \(D(x)\) represents the output of the discriminator and \(E\) is the expectation value (average). \(P_r\) and \(P_g\) are the real and generated data distributions, respectively. And \(P_d\) is the ‘sampling distribution’ which samples data from both the real and generated distributions, as discussed in [16]. The value \(\lambda = 10\) was used as recommended as well. Then the parameters are updated with an optimizer,

\[
w \leftarrow w + \alpha \cdot \text{Optimizer}(w, \nabla_w \text{Loss}_D),
\]

\[
\theta \leftarrow \theta - \alpha \cdot \text{Optimizer}(\theta, \nabla_\theta \text{Loss}_G),
\]

where \(w\) are the discriminator’s parameters, \(\theta\) is the generator’s parameters, and \(\alpha\) is the learning rate. RMSProp was chosen as the optimizer [17], after testing out several other options, such as Adam [18]. The architecture of ScGAN is similar to the one used in [10]. Since it was shown that the model worked well for non-superconducting, inorganic, stoichiometric compounds, similar layer design was used for a model for superconductors.

2.3.1. Training. The model was first trained on the OQMD dataset for 400 epochs. It was then transfer learned onto the SuperCon dataset or a subset of it, on which it would train for another 500 epochs. Transfer learning onto four different datasets (cuprates, pnictides, others, and everything together) resulted in four different versions of the GAN. Afterwards, the testing procedure and the data analysis were conducted, and then the hyperparameters were updated based on the results. This updating was done manually: the hyperparameters such as compound data encoding, layer sizing, etc were adjusted after training. Note that the same hyperparameters were used for the four different networks. One could individually modify the parameters for each, but the point is that this one architecture can learn any of the superconducting classes. The training curves for the final model on each of these sets are displayed in figure 5. In some cases the training was stopped early to prevent the model from overfitting.

![Figure 4](image-url)
Figure 5. The generator loss against training epoch for each of the four datasets the GAN trained on: (a) all of SuperCon; (b) others (i.e. not cuprates or pnictides); (c) cuprates; and (d) pnictides.

We note that more analysis could be done in the future to improve the learning process. For example, there is evidence of overfitting (in panel (b)), and the hyperparameters can be adjusted further to improve the model’s performance. Furthermore, some quantitative analysis could be done to look at the impact of transfer learning and how it affects the quality of the final model. However, the model presented here still shows great promise despite these shortcomings.

2.3.2. Testing. After each training process, 5000 hypothetical compounds were generated from the versions of the GAN trained on the smaller sets and 30000 from the version trained on everything. The generated predictions were then inspected with various quality checks. Each compound was first tested for validity using the centerge neutrality and electronegativity check features of the SMACT package [19]. The package tests the compound for electronegativity and centerge balance to determine whether it is a valid compound or not. Each prediction was then checked for uniqueness—whether the compound showed up earlier in the generated list—and novelty—whether the compound was in the training dataset. These three checks looked at the general quality of the model. Then, more specific to superconductivity, each compound was run through the model from [6] to check whether it is a superconductor or not, as well as to predict its critical temperature. Lastly, the formation energy of each generated compound was calculated using the model from [20], which indicates the stability of the compound. These tests will be discussed further in section 3, along with the actual results.

2.4. Clustering

In order to further assess the quality of predictions, clustering analysis was performed. Clustering is an unsupervised machine learning technique whose main goal is to unveil hidden patterns in the data. It was recently applied to superconductors from SuperCon database [15]. Depending on the data set, different clustering algorithms were used, such as $k$-means, hierarchical, Gaussian mixtures, self-organizing maps, etc. The results showed that in case of superconductors, clustering methods can achieve, and in some cases exceed, human level performance.

In order to visualize clustering results, different techniques can be used. It was shown that for superconductors the so-called t-SNE produces the best results [15]. t-SNE is a non-linear dimensionality reduction technique which allows higher dimensional data to be represented in 2D or 3D [21]. In case of superconductors, the data points are 96-dimensional (each compound is represented by a $1 \times 96$ matrix), as discussed in section 2.2. t-SNE reduces these dimensions down to either two or three, which allows easy visualization. We point out, however, that these reduced dimensions do not have any physical meaning.
Table 2. The percentage of generated predictions that were novel (not in the training set) and unique (distinct from others in the given generated set) for each of the versions of ScGAN trained on the given training datasets on the left.

| GAN version         | Novel % | Unique % |
|---------------------|---------|----------|
| Entirety of SuperCon| 99.69%  | 96.78%   |
| Cuprate             | 99.74%  | 92.98%   |
| Pnictides           | 99.32%  | 58.74%   |
| Others              | 98.89%  | 91.58%   |

Table 3. The counted versions of the table above.

| GAN version         | Distinct generations | Distinct, novel generations |
|---------------------|----------------------|----------------------------|
| Entirety of SuperCon| 29034                | 28944                      |
| Cuprate             | 4637                 | 4649                       |
| Pnictides           | 2937                 | 2917                       |
| Others              | 4579                 | 4528                       |

3. Results

After training, from the four different versions of the GAN, we generated four superconductor candidate lists with either 5000 or 30000 chemical compositions. We then ran the predicted superconductors through a series of tests to evaluate their quality. The first few were general tests, and the rest were in the context of superconductivity using existing computational models, as experimentally testing all of them is unfeasible.

3.1. Duplicates and novelty

We first screened the output for duplicates within the generated sets themselves and then for duplicates between the generated set and the dataset of known superconductors that it trained on. The results are tabulated in tables 2 and 3. As seen in the table, the number of repeats within the generated samples were relatively low (high uniqueness), with the exception of pnictides, which had more duplicates than the rest. This is likely due to the fact that the dataset of pnictides was significantly smaller than the rest (see table 1). We speculate that this overall low rate of duplicates stems from the fact that the model is able to handle decimals (see section 2.2 and figure 4), which opens up a large composition space for it to explore.

The percent of predicted compounds that were novel, i.e. not already known to exist, listed in table 2 is also very high across all four GANs. These two results demonstrate that all versions of ScGAN can generate a diverse array of new compositions.

However, note that these duplicates were not discounted from the following evaluations because if they are still tested to be superconducting then they are still valid generations from the model, regardless of whether or not they duplicate something from the training set. These prior tests were simply to show how much the model is able to give us new predictions, which are the valuable ones.

3.2. Formation energy

As mentioned in the previous section, we found the formation energies of the predicted compounds from the GANs using ElenNet [20]. It was indicated in [20] that a negative value of formation energy is a good indicator of stability, i.e. the possibility of being synthesized in the lab. In figure 6 we display the values of formation energy for all predictions. We see from the distributions of formation energies that most of predicted compounds have calculated formation energies less than zero. Even though this does not provide definitive proof of stability, it is a general indication that most of the predicted compounds are stable and can be synthesized in the lab.

3.3. Superconductivity

As a next test, we ran the predicted compounds through the KNNs classification model from [6] for predicting superconductivity based on elemental composition, in order to check if the predictions of two machine learning models (GAN and KNN) would agree. However, the probabilistic nature of the machine learning model had to be taken into account. If $p_{sc}$ is the proportion of predictions that came up superconducting according to the model, then we can estimate $p_{sc}$, the true proportion of superconducting entries, using Bayesian statistics. Denoting $tp$ and $fp$ as the true positive and false positive rates of the classification model, respectively, we can write

$$\rho_{sc} \cdot tp + (1 - \rho_{sc}) \cdot fp \approx p_{sc}. \quad (6)$$

Solving for $\rho_{sc}$ gives

$$\rho_{sc} \approx \frac{p_{sc} - fp}{tp - fp}. \quad (7)$$

The true positive and false positive rates here are reported from [6]: $tp = 98.69\%$ and $fp = 16.94\%$. The output percentages along with the estimates of the true proportions calculated from equation (7) are tabulated in table 4.

All four GANs achieved very high percentages of superconducting material according to the KNN model, especially when compared to the 3% figure from the manual search in [3]. However, the only definite test of superconductivity can come from experimental measurements.

3.4. Critical temperature estimates

We also calculated the critical temperatures of our predictions using the regression model from [6]. Similar to the superconductivity tests in the previous subsection, these calculated values can only be taken as approximations. However, the regression model can still provide us with a general understanding of the capabilities of ScGAN. The critical temperature outputs from the model are summarized in table 5 and the distributions are in shown in figure 7. While the distributions are somewhat broad, the GANs were still able to find several superconductors with predicted critical temperatures higher than 100K, which exceeds the manual search maximum of 58K
(though that search was mostly restricted to pnictides) and several of the previous machine learning approaches [5, 6]. This is not surprising, as those previous searches were limited to existing databases of stoichiometric compounds, which meant that these forward design approaches could only produce a limited number of candidates, mostly with low critical temperatures.

### 3.5. Ability to learn features

We then looked at the types of superconductors that were generated by the different versions of ScGAN. The distributions are given in table 6, and we can see that each versions of ScGAN generated mostly superconductors that matched their training data. This indicates the GAN was able to detect the different underlying features behind these different major classes of superconductors.

### 3.6. Clustering results

In figure 8 we display the results of clustering analysis on three sets of predictions: panel (a) for cuprates, panel (b) for pnictides and panel (c) for others. The results are visualized with the help of t-SNE. As pointed out above, the two t-SNE dimensions, $Y_1$ and $Y_2$, do not have any physical meaning. Full circles of different colors represent different families of superconductors from SuperCon database, whereas purple open circles represent GAN predictions. As can be seen from all three panels, GANs were able to generate new superconductors from all known families of cuprates, pnictides and other superconductors. However, GANs did not predict any new families of superconductors.

### 3.7. Promising candidates

After running through the $T_c$ prediction model, we manually identified the ones that looked the most promising, including some with very high critical temperatures. It turns out that most of these were cuprates, which is not surprising.
Figure 7. Distributions of the critical temperatures of the predictions of the four different versions of ScGAN: trained on (a) everything, (b) others, (c) cuprates, (d) pnictides.

Table 6. The distribution of the predicted superconductors across the different classes of superconductors for the different versions of the GAN.

| Training data | Cuprate % | Pnictide % | Other % |
|---------------|-----------|------------|---------|
| Cuprate       | 92.76%    | 0.06%      | 7.18%   |
| Pnictides     | 0.02%     | 99.84%     | 0.14%   |
| Others        | 0.14%     | 0.6%       | 99.26%  |

Figure 8. Clustering of the predicted compounds from various versions of the GAN: (a) cuprates, (b) pnictides and (c) others. Full circles represent the data points from SuperCon and purple open circles are GAN predictions.

Considering that superconductors with highest critical temperatures in the training set are cuprates. We then checked the Crystallography Open Database (COD) [22] to see if those compounds were in the database. The ones listed in table 7 could not be found neither in COD nor in SuperCon, showing that this model overcomes the limitations of the
previous forward design models and finds completely new superconductors. A more comprehensive list of predictions is available upon reasonable request.

4. Conclusion

For decades the search for new superconductors has relied on the serendipity of material scientists to synthesize a new material with superconducting proprieties. This paper introduced a novel method to search for superconductors—discovering candidates with a GAN. In contrast to previous computational methods which attempted to predict new superconductors off of existing datasets, this model predicted compounds directly. This ‘inverse design’ approach proved to be far more powerful than manual search methods and previous computational methods, with the model being able to generate thousands of candidates with a wide-range of critical temperatures that lied outside of existing databases (both superconductor and general inorganic compound databases). Even while only training on chemical compositions, more than 70% of the GANs predictions were cross-checked with a separate model to be potentially superconducting (the only way to know for sure, however, would be to synthesize and check these compounds in a lab). Of these, several were promising HTS candidates listed in table 7. We point out that previous models would have been unable to find such candidates as they were outside of current databases.

While the compounds generated were new, our clustering showed that the GAN did not generate any new families of superconductors. However, it was still able to generate non-stoichiometric compounds, widening the scope of the computational search. Future studies should look into some improvements that can be made, such as being able to account for centege neutrality and crystal structure in the compound encodings. Chemical checks on the predicted compounds (using SMACT as detailed earlier) revealed that while there was electronegativity balance, centege balance was not always exact for the superconductor candidates. Furthermore, crystal structure has been known to play a significant role in superconductivity, so including it in calculations would most likely result in improvements. However, this may prove to be a difficult endeavor as crystal structure is not well-documented in existing databases. Active transfer learning could also be attempted to narrow down the predictions to HTSs only [23]. However, even without these possible improvements, the model at its current version is still very promising and can be applied to search for superconductors, starting with the candidates identified in this paper.

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID IDs

Evan Kim https://orcid.org/0000-0001-9764-1493
S V Dordevic https://orcid.org/0000-0001-8360-7577

References

[1] Arute F et al 2019 Quantum supremacy using a programmable superconducting processor Nature 574 505–10
[2] Hirsch J E, Maple M B and Marsiglio F 2015 Superconducting materials classes: introduction and overview Physica C 514 1–8
[3] Hosono H, Tanabe K, Takayama-Muromachi E, Kageyama H, Yamanaka S, Kumakura H, Nohara M, Hiramatsu H and Fujitsu S 2015 Exploration of new superconductors and functional materials and fabrication of superconducting tapes and wires of iron pnictides Sci. Technol. Adv. Mater. 16 033503
[4] Bedolla E, Padierna L C and Castañeda-Priego R 2020 Machine learning for condensed matter physics J. Phys.: Condens. Matter 33 053001
[5] Stanev V, Oses C, Kusne A G, Rodriguez E, Paglione J, Curtarolo S and Takeuchi I 2018 Machine learning modeling of superconducting critical temperature npj Comput. Mater. 4 29
[6] Rother B and Dordevic S V 2020 Predicting new superconductors and their critical temperatures using machine learning Physica C 575 1353689
[7] Konno T, Kurokawa H, Nabeshima F, Sakishita Y, Ogawa R, Hosako I and Maeda A 2021 Deep learning model for finding new superconductors Phys. Rev. B 103 014509
[8] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets Proc. 27th Int. Conf. on Neural Information Processing Systems (NIPS’ 14) vol 2 (MIT Press) pp 2672–80
[9] Dash A, Ye J and Wang G 2021 A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines—from medical to remote sensing (arXiv:2110.01442)

[10] Dan Y, Zhao Y, Li X, Li S, Hu M and Hu J 2020 Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials npj Comput. Mater. 6 84

[11] Hu T, Song H, Jiang T and Li S 2020 Learning representations of inorganic materials from generative adversarial networks Symmetry 12 1889

[12] National Institute for Materials Science 2023 SuperCon database (available at: https://mdr.nims.go.jp/)

[13] 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

[14] 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

[15] Roter B, Ninkovic N and Dordevic S V 2022 Clustering superconductors using unsupervised machine learning Physica  C 598 1354078

[16] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C 2017 Improved training of Wasserstein GANs Advances in Neural Information Processing Systems vol 30

[17] Hinton G 2012 Neural networks for machine learning: lecture 6a overview of min-batch gradient descent (available at: www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf)

[18] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[19] Davies D W, Butler K T, Jackson A J, Morris A, Frost J M, Skelton J M and Walsh A 2016 Computational screening of all stoichiometric inorganic materials Chem 1 617–27

[20] Jha D, Ward L, Paul A, Liao W-K, Choudhary A, Wolverton C and Agrawal A 2018 ElemNet: deep learning the chemistry of materials from only elemental composition Sci. Rep. 8 17593

[21] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE J. Mach. Learn. Res. 9 2579–605

[22] Vaitkus A, Merkys A and Gražulis S 2021 Validation of the crystallography open database using the crystallographic information framework J. Appl. Crystallogr. 54 661–72

[23] Kim Y, Kim Y, Yang C, Park K, Gu G X and Ryu S 2021 Deep learning framework for material design space exploration using active transfer learning and data augmentation npj Comput. Mater. 7 140