npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials

Yabo Dan[1], Yong Zhao[2], Xiang Li[1], Shaobo Li[1,3], Ming Hu[4] and Jianjun Hu[1,2]✉

A major challenge in materials design is how to efficiently search the vast chemical design space to find the materials with desired properties. One effective strategy is to develop sampling algorithms that can exploit both explicit chemical knowledge and implicit composition rules embodied in the large materials database. Here, we propose a generative machine learning model (MatGAN) based on a generative adversarial network (GAN) for efficient generation of new hypothetical inorganic materials. Trained with materials from the ICSD database, our GAN model can generate hypothetical materials not existing in the training dataset, reaching a novelty of 92.53% when generating 2 million samples. The percentage of chemically valid (charge-neutral and electronegativity-balanced) samples out of all generated ones reaches 84.5% when generated by our GAN trained with such samples screened from ICSD, even though no such chemical rules are explicitly enforced in our GAN model, indicating its capability to learn implicit chemical composition rules to form compounds. Our algorithm is expected to be used to greatly expand the range of the design space for inverse design and large-scale computational screening of inorganic materials.

npj Computational Materials (2020)6:84 ; https://doi.org/10.1038/s41524-020-00352-0

INTRODUCTION

Discovering new inorganic materials such as solid electrolytes for lithium-ion batteries is fundamental to many industrial applications. While recent years have observed tremendous efforts on rational materials design, progress has been limited due to the challenge to find new materials that meet diverse technical and economic constraints. From the computational perspective, brute-force molecular simulations or first-principles methods are computationally too expensive for large-scale screening of the vast chemical space. A recent effort[1] to quantify the magnitude of the compositional space for multi-component inorganic materials showed that even after the application of chemical filters such as charge neutrality or electronegativity balance, the space for four-component/element materials exceeds 10^{10} combinations and the five-component/element space exceeds 10^{13} combinations. Indeed, a machine learning (ML) based model has been applied to screen billions of hypothetical materials to identify promising high ion-conductors[2]. Considering the huge space of doped materials with different mixing ratios of elements and many applications such as high-temperature superconductors, where six to seven component materials are common, the number of potential materials is immense. Such combinatorial explosion calls for the need for more effective sampling approaches to search the chemical design space that employ existing explicit chemical and physical knowledge and also implicit elemental composition knowledge embodied within known synthesized materials. To gain more efficient search, a variety of explicit chemical rules for assessing the feasibility of a given stoichiometry and the likelihood of particular crystal arrangements have been used in computational screening such as the Pauling's rules (charge neutrality), electronegativity balance, the radius ratio rules[3], Pettifor maps[4] and etc. However, such approaches still fail to capture enough implicit chemical rules to achieve efficient chemical design space sampling.

Recently, generative machine learning models such as autoencoders (AE) and its variants (VAE, AAE), RNNs, generative adversarial networks (GANs) have been successfully applied to inverse design of organic materials[5–8]. These algorithms mainly exploit the sequential or graph representations of organic materials to learn the composition rules of the building blocks for generating valid and novel hypothetical materials. Given a large set of samples, a GAN is capable of learning complicated hidden rules that generate the training data, and then applies these learned rules to create new samples with target properties. When applied to inverse design, GANs have demonstrated their power in efficient sampling of design space[5,9], more efficient than other sampling approaches such as random sampling[10], Monte Carlo sampling, and other heuristic sampling (such as genetic algorithms[11]). However, due to the radical difference in building blocks and their composition rules, such generative machine learning models have not been applied to the generation of inorganic materials so far to the best of our knowledge. Recently, variational autoencoders[11,12] have been proposed to generate hypothetical crystal structures of inorganic materials. However, these methods are either limited to generate new structures of a given material system such as the V–O system[11] or tend to mostly generate crystal materials that are not physically stable[12]. A GAN model for discovering new crystal structures have also been recently proposed[13]. However, their study is limited to very restricted binary systems and lacks composition diversity for generated structures. Another related work[14] proposes to use conditional GAN and VAE models for composition generation. However, they use real-valued one-hot encoding of materials, fully connected neural network architecture, and mixed Wasserstein

[1]School of Mechanical Engineering, Guizhou University, Guiyang 550025, China. [2]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201, USA. [3]Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang 550025, China. [4]Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29201, USA. ✉email: jianjunh@cse.sc.edu

losses and least square loss of properties, which makes their models tend to generate invalid compositions with very low composition diversity.

In this paper, we propose the first generative adversarial network model for efficient sampling of inorganic materials design space by generating hypothetical inorganic materials. Trained with materials from inorganic materials databases such as OQMD[15], Materials Project[16], and ICSD, our GAN models are able to learn the implicit chemical compositional rules from the known materials to generate hypothetical but chemically sound compounds. Without explicitly specifying the chemical rules, our GANs trained with all charge-neutral and electronegativity-balanced samples of the ICSD subset can generate hypothetical materials with 84.5% reproducing the charge neutrality and balanced electronegativity. The analysis shows that our generative GAN can achieve much higher efficiency in sampling the chemical composition space of inorganic materials than the exhaustive enumeration approach.

RESULTS

Representation of inorganic materials

Through simple statistical calculation of the materials in the OQMD dataset[17], 85 elements are found, and each element usually has <8 atoms in any specific compound/formula. We then represent each material as a sparse matrix $T \in R^{d \times s}(d = 8,\ s = 85)$ with 0/1 integer values. Each column represents one of the 85 elements while the column vector is a one-hot encoding of the number of atoms of that specific element (see Supplementary Fig. 5). The binary one-hot encoding is chosen as two of the major chemical rules for valid chemical compositions including the charge neutrality and electronegativity balance are checked by the discrete atom numbers of composing elements and our binary encoding allows us to take advantage of the strong binary pattern learning capability of convolutional neural networks used in our GAN models. Here the columns (elements) are sorted by their atomic number and rows are sorted from 0 to 7.

The GAN generation model

Generative models can be built on several machine learning algorithms such as variational autoencoder (VAE), generative adversarial networks (GAN), reinforcement learning (RL), recurrent neural networks (RNN), and their hybrids[5]. Different from other generative models[18,19], GANs do not directly use the discrepancy of the data and an assumed model distribution (as VAE) to train the generator. Instead, it uses an adversarial training approach: it first trains a discriminator to differentiate real samples from faked samples, which then guides the training of the generator to reduce this difference. These two training processes are alternatively repeated. Their arm race will lead to high performance of both the generator and the discriminator.

Our generative ML model for inorganic materials (MatGAN) is based on the GAN scheme as shown in Fig. 1.

We choose the 8 × 85 matrix representation of materials samples to build the GAN model. We found the integer representation of materials greatly facilities the GAN training. In our GAN model, both the discriminator (D) and the generator (G) are modeled as a deep neural network. The generator is composed of one fully connected layer and seven deconvolution layers. The discriminator is composed of seven convolution layers followed by a fully connected layer. Each of the convolution and deconvolution layer comes with a batch normalization layer. The output layer of the generator uses the Sigmoid function as the activation function while all other batch normalization layers use the ReLu as the activation function. The detailed network configuration is shown in Supplementary Table 2. In order to avoid the gradient vanishing issue of standard GANs, we adopt the Wasserstein GAN[20], which replaces the JS divergence distance with the Wasserstein distance. The GAN model will be trained using the Wasserstein GAN approach by minimizing both the generator loss and discriminator loss, which are defined as

$$\mathrm{Loss_G} = -\mathrm{E}_{x:P_g}[f_w(x)] \tag{1}$$

$$\mathrm{Loss_D} = \mathrm{E}_{x:P_g}[f_w(x)] - \mathrm{E}_{x:p_r}[f_w(x)] \tag{2}$$

where, P_g and P_r are the distributions of generated samples and real samples; $f_w(x)$ is the discriminant network. Equations (1) and (2) are used to guide the training process. The smaller the $\mathrm{Loss_D}$, the smaller the Wasserstein distance between the generated samples and the real samples and the better the GAN is trained.

Variational autoencoder for evaluating GAN performance

During our GAN generation experiments for OQMD dataset, we found that it sometimes has difficulty to generate a specific category of materials. This may be caused by the limited samples to learn the required composition rules to generate those samples. To investigate this issue, we built an autoencoder (AE)[21] model as shown in Fig. 2. The autoencoder is composed of an encoder with seven convolutional layers followed by a fully connected layer and a decoder composed of a fully connected layer followed by seven deconvolution layers. After each of the convolution and deconvolution layer, there is a batch normalization layer used to speed up

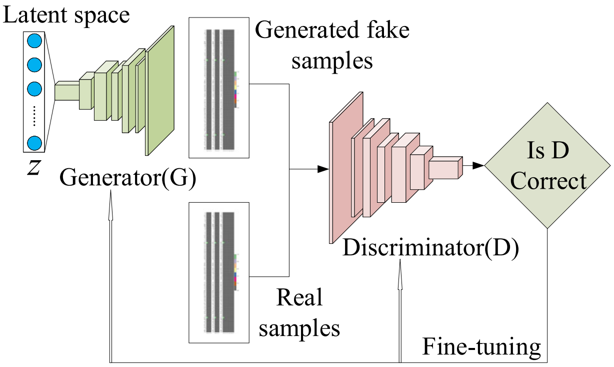

Fig. 1 Architecture of MatGAN for inorganic materials. It is composed of a generator, which maps random vectors into generated samples and a discriminator, which tries to differentiate real materials and generated ones. Detailed configuration parameters are listed in Supplementary Table 1 and Supplementary Fig. 1.

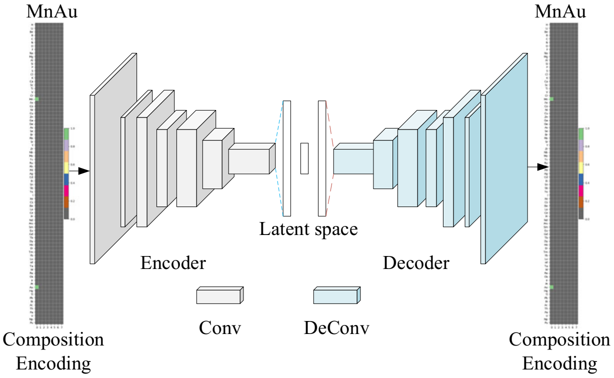

Fig. 2 Architecture of autoencoder. Detail configuration parameters are shown in Supplementary Table 2 and Supplementary Fig. 2.

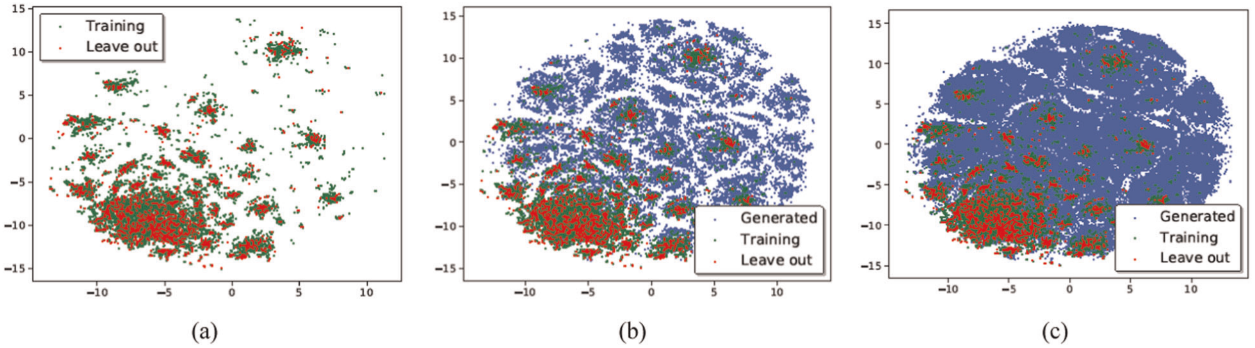

Fig. 3 Inorganic materials space composed of existing ICSD materials and hypothetical materials generated by GAN-ICSD. The two axes correspond to the two dimensions after t-sne-based dimension reduction. The ICSD materials only occupies a tiny portion of the chemical space of inorganic materials. **a** Training samples (green dots) and leave-out validation samples (red dots) from ICSD; **b** 50,000 generated samples (blue dots) together with training and leave-out samples; **c** 200,000 generated samples together with training and leave-out samples.

training and reduce the influence of initial network weights[22]. The ReLu is used as the activation function for all the batch normalization layers. The Sigmoid function is used as the activation function for the decoder's output layer. The detailed configuration parameters are listed in Supplementary Table 2. The autoencoder are trained with 291,840 inorganic materials selected from the OQMD database. In order to ensure the overlap between the original input matrix T and the matrix reconstructed by the decoder as much as possible, we adopt the negative dice coefficient[23] commonly used in medical image semantic segmentation as the loss function of AE. The AE model is then trained using the back-propagation algorithm. The loss function is shown in the following equation:

$$\mathrm{Loss}_{\mathrm{AE}} = -\mathrm{Dice} = -\frac{2|A \cap B|}{|A| + |B|} \approx -\frac{2 \times A \cdot B}{\mathrm{Sum}(A) + \mathrm{Sum}(B)} \quad (3)$$

where $A \cap B$ denotes the common elements of A and B, |g| represents the number of elements in a matrix, · denotes dot product, Sum(g) is the sum of all matrix elements. Dice coefficient essentially measures the overlap of two matrix samples, with values ranging from 0 to 1 with 1 indicating perfect overlap[24].

The decoder module of the AE model shares the same architecture of the generator in our MatGAN model. Our hypothesis is that if the trained AE model cannot decode a specific material, it is unlikely our GAN model can generate it. By screening out the non-decodable materials out of the OQMD database using the AE, we may obtain a deeper understanding of the limitations of our GAN models.

The model performance of GANs

Efficient sampling of the inorganic chemical space by the GANs. We trained our GANs according to the procedures as detailed in the "Methods" section. For all these GANs, including GAN-OQMD, GAN-MP, and GAN-ICSD, we then generated 2 million hypothetical materials using each of these generators and evaluate their validity, uniqueness, and novelty.

Mapping inorganic materials design space

Out of the 2 million samples generated by the GAN-ICSD, we filter out all samples that do not satisfy charge neutrality and balanced electronegativity leading to 1.69 million generated samples. To visualize how the generated ones are distributed compared to the training datasets from ICSD, we applied T-sne dimension reduction technique[25] to reduce the dimension of the matrix representations of the samples from the generated set, the training set, and the leave-out validation set. The distribution of the generated samples versus the training and validation set are shown in Fig. 3. It is observed that the training samples from ICSD occupy only a very small portion of the whole space. The GAN-ICSD, however, has been able to generate potentially interesting hypothetical materials that fill the design space, which may significantly expand the range of the ICSD database.

Validity check. Charge neutrality and electronegativity balance are two fundamental chemical rules of crystals. It is thus interesting to check how the generated samples from our GAN models satisfy these rules without explicit enforcement of such rules during model training. To do this, we adopt the charge-neutrality and electronegativity check procedure as proposed in ref. [1] to calculate the percentages of samples that obey these rules within the training and generated sets of all four databases. The results are shown in Fig. 4. First, we found that the percentages of the validly generated samples are very close to those of the training set. For OQMD, when the training set has 55.8% charge-neutral samples, the generated set has 56.1%. For MP and ICSD, the percentage of generated charge-neutral samples (84.8 and 80.3%) are also close to those of the training sets (83.5% and 84.4%). Similar observations are found for electronegativity check. It is impressive that when we ensure all training samples in the ICSD_filter are charge-neutral and electronegativity-balanced, up to 92.1% and 84.5% of the generated samples satisfy the two chemical rules, respectively, despite that no such rules are explicitly modeled or enforced in our GAN training models. To demonstrate the significance of this high percentage of chemically valid candidates, we compare our results to the exhaustive enumeration approach in ref. [1], Table 1. The percentage of all binary/ternary/quaternary samples that satisfy both charge neutrality and electronegativity is 0.78% with exhaustive enumeration compared with our 62.24%, which corresponds to 77 times of enrichment in terms of sampling efficiency. This strongly indicates that our GAN models have successfully learned implicit chemical rules for generating chemically valid hypothetical materials.

Formation energy distribution of generated materials

another way to evaluate the quality of generated hypothetical materials is to check their stability, which can be measured by their formation energy[26]. To do this, first, GAN-OQMD, GAN-ICSD, and GAN-MP were used to generate 2 million materials candidates each. Then, we selected all the materials with lithium element and then filter out all those materials that do not satisfy charge neutrality and balanced electronegativity. Finally, we obtained 15,591, 137,948, and 281,320 lithium-containing compounds, respectively, from GAN-OQMD, GAN-ICSD, and GAN-MP. We then downloaded the formation energy prediction machine learning model (ElemNet) developed by Jha et al.[26] and then used it to

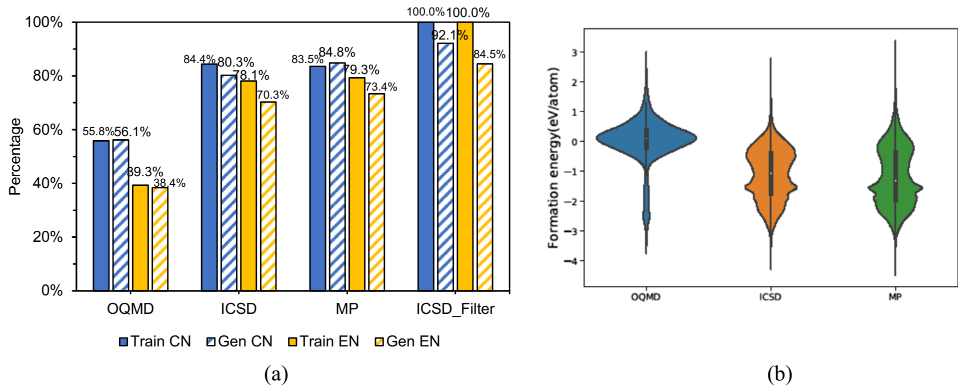

Fig. 4 Evaluation of the validity of generated materials. a The percentages of charge-neutral (CN) and electronegativity-balanced (EN) samples of the generated samples are very close to those of the training sets for all four datasets. Train/gen CN: percentage of training/generated samples that satisfy charge neutrality; Train/gen EN: percentage of samples that satisfy balanced electronegativity. **b** Formation energy distribution of the Li-containing compounds generated by three GANs. Both GAN-ICSD and GAN-MP can generate a large percentage of hypothetical materials with low (<0) formation energy.

Table 1. Novelty check of generated samples by GANs.

	GAN-OQMD	GAN-MP	GAN-ICSD
Training sample #	251,368	57,530	25,323
Leave-out sample #	27,929	6392	2813
Generated sample #	2,000,000	2,000,000	2,000,000
Recovery % of training samples	60.26%	47.36%	59.54%
Recovery % of leave-out sample	60.43%	48.82%	60.13%
New samples	1,831,648	1,969,633	1,983,231

predict the formation energies of all these hypothetical materials. Figure 4b shows that the formation energy of these generated materials is mostly <0, especially for those generated by GAN-ICSD and GAN-MP, which are trained with more chemically valid samples. Also, much higher percentage of generated samples by the GAN-OQMD are found to have higher formation energy scores in the figure, which is due to the fact that 68.48% training samples of OQMD have formation energies larger than 0.

Uniqueness check. To check the uniqueness of the generated samples, we calculate the percentages of the number of unique samples out of the number of all generated samples (n) as n goes from 1 to 340,000 for all three GANs trained on the OQMD, MP, and ICSD datasets, respectively (Fig. 5a). First, it can be found that with the generation of more and more samples, the percentage of unique samples goes down, showing that it is more difficult to generate new hypothetical materials. However, even after generating 340,000 samples, our GANs still maintain a 68.09%, 85.90%, and 73.06% uniqueness for GAN-OQMD, GAN-MP, and GAN-ICSD, respectively. While all three curves decay with increasing number of generated samples, the GAN-MP maintains higher percentage of unique samples. Actually, the uniqueness curve of GAN-MP dominates the one of GAN-ICSD, which further dominates the one of GAN-OQMD. After close examination of the distributions of training and generated samples in terms of their element numbers, we found that this is mainly due to the distribution bias of the training sets of the three GANs (see Supplementary Fig. 3). For GAN-OQMD, the training set is dominated by ternary compounds (84.4%) and it tends to generate ternary samples while the total number of chemically valid materials as estimated by SMACT[1] (Semiconducting Materials from Analogy and Chemical Theory) to be around 200,000. So, it tends to generate many duplicate ternary samples. For GAN-ICSD, the ratio of binary/ternary/quaternary is about 2:3:1, which allows it to generate more diverse samples, leading to higher uniqueness curve. For GAN-MP, the ratio of binary/ternary/quaternary is about 0.8:2:1, which is much more balanced than those of GAN-OQMD and GAN-ICSD and it also has much more quaternary and quinary training samples (see Supplementary Table 3). This allows it to generate most diverse samples.

Novelty check. To check the capability of our GANs to generate novel materials, we use the hold-out validation approach. We first leave out 10% samples from each of the three datasets OQMD, MP, and ICSD. Then we train the GANs and use them to generate a certain number of samples. We then examine what percentage of training samples and hold-out validation samples have been recovered/re-discovered and how many new samples have been generated. The results are shown in Table 1. First, we found that when the GANs recover/generate a certain percentage of training samples, the approximate corresponding percentages of validation (hold-out novel) samples are also recovered. For example, when the GAN-MP recovered 47.36% of its training set, about 48.82% of the hold-out samples have also been simultaneously generated. This demonstrates that our GANs can be used to discover new materials that do not exist in the training set. To further understand the generation performance, we calculated the recovery percentages of the training set and the leave-out validation set along with the percentages of new samples for binary, ternary, and quaternary samples (Fig. 5b). First, by generating 2 million samples, GAN-ICSD has generated 78.1% training binary samples while also generating/rediscovering 82.7% leave-out validation binary materials. The recovery rates drop to 30.4% and 31.2%, respectively, or ternary training and validation samples as the number of possible ternary samples are larger than binary ones, which also explains the recovery rates dropping to 3.3 and 5.2% for quaternary training and validation sets. In addition, out of all the generated binary/ternary/quaternary samples, 83.15%/98.68%/99.98% of them are novel hypothetical materials, which strongly shows the capability of our GAN model to generate new materials candidates as a majority of these new candidates satisfy the basic chemical rules as shown in Fig. 4.

Conditional generation of hypothetical materials by GAN. In addition to generating valid inorganic materials, it is interesting to check if our GAN models can generate new materials with desired properties by sampling from the generative distribution estimated by the model[27]. To verify this, we collected 30,186 inorganic materials from Materials Project whose bandgap values are larger than 0. We then use these high-bandgap materials set to train a GAN-bandgap model aiming to generate hypothetical

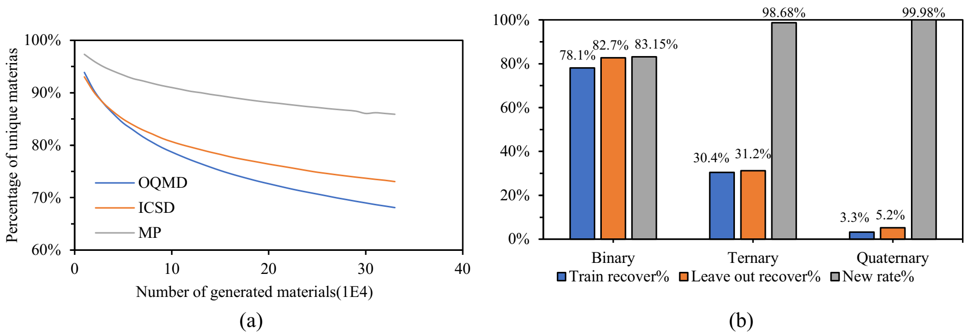

Fig. 5 Uniqueness and novelty check of the generated materials. a Comparison of uniqueness curves of the hypothetical materials generated by three GANs. GAN-MP achieves the dominating curve due to its more balanced distribution of binary/ternary/quaternary training samples. **b** Distribution of recovery rates of training and validation samples, and also percentages of new generated hypothetical materials.

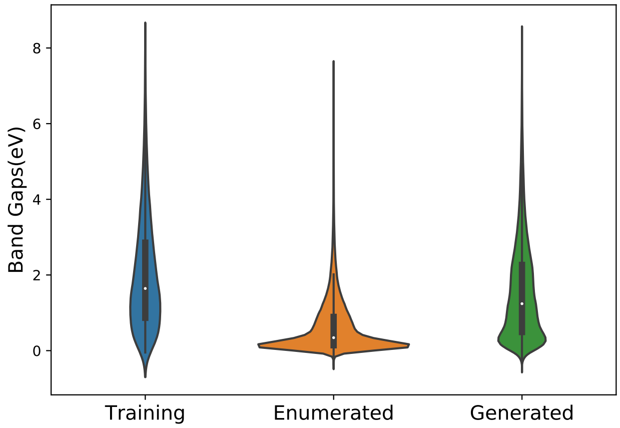

Fig. 6 Comparison of bandgap distributions of the generated materials by GAN-bandgap, the training set, and the enumerated set. The GAN-generated samples show much more similar bandgap distribution than the enumerated compositions.

high-bandgap materials. To verify the bandgap values of generated samples, we trained a bandgap prediction model using the Gradient-Boosted Decision Tree (GBDT) machine learning algorithm with Magpie features[28] (see the Methods section for its training details). We also use this model to predict the bandgap values of the exhaustively enumerated materials set. Figure 6 shows the distribution of the band gaps of the generated materials set versus those of the training set and the exhaustively enumerated set. The bandgap distribution of generated samples is much similar to that of the training set, which demonstrates the capability of our GAN-bandgap can generate hypothetic high-bandgap materials efficiently.

Discovery of potential new materials. To evaluate how likely our GAN models can generate confirmed new materials, we take a cross-validation evaluation approach. Essentially, for all the new hypothetical materials generated by each of our GAN models, we check how many of them are confirmed/included by the other two datasets. Table 2 lists the cross-validation confirmation results. It is found that out of the 2 million generated materials by GAN-ICSD, 13,126 materials are confirmed by and included in the MP dataset and 2349 new materials are confirmed by the OQMD dataset. GAN-MP also has 6880 and 3601 generated samples confirmed by ICSD and OQMD, respectively. Another validation of the predictive power of our GAN models come from a recent study on discovering four new materials of aluminum and sulfide using DFT-based crystal structure prediction software CALYPSO[29]. Two of the four materials are in our candidate list: AlS_2, AlS, which are confirmed by their study and we have predicted an additional undiscovered AlS_3.

Table 2. Cross-validation confirmation of generated new materials by our GANs.

	ICSD dataset	MP dataset	OQMD dataset
GAN-ICSD	N/A	13,126	2349
GAN-MP	6880	N/A	3601
GAN-OQMD	3428	58,603	N/A

Limitation of MatGAN examined by autoencoder. Here we aim to check the relation of AE non-decodable materials and the difficulty of our GANs to generate them. To train the AE model, we randomly split the OQMD_L dataset with 90% samples for AE training and 10% samples as testing. The learning rate is set as 10^{-3}, batch size 1024, and Adam optimizer is used. The final AE model is picked as the model with the best performance over the test set within 1000 epochs of training. We found that our AE model can decode 96.31% and 95.50% of the samples from the training set and the test set. These samples seem to share some common chemical composition rules.

To show the difference between the decodable samples and non-decodable ones, we applied T-sne dimension reduction technique[25] to reduce the dimension of the matrix representations of all OQMD_L dataset to 2 and then visualize 20% of the samples on the 2D plot (Fig. 7), in which the red dots represent non-decodable samples while blue ones represent decodable ones. The apparent different distributions show that these two categories of samples have different composition rules. Our hypothesis is that the decodable samples share well-established chemical composition rules, which allows our GAN generators for efficient sampling of the corresponding chemical space. On the other hand, the non-decodable samples will be difficult to generate by our GAN model. To verify this, we calculated the percentage of non-decodable samples that have been generated by the trained GAN-OQMD. It is observed that almost 95% of the non-decodable materials are out of the scope of the generated samples even after generating 2 million of samples while 60.26% of the decodable training samples have been re-discovered.

This shows that our GANs have limitation in generating non-decodable materials type. It also means that non-decodable materials have special composition rules that either need more data or more powerful generator models to learn. Indeed, comparison on the enriched element distribution analysis (Supplementary Fig. 4) shows that the decodable and non-decodable materials have distinct element distributions.

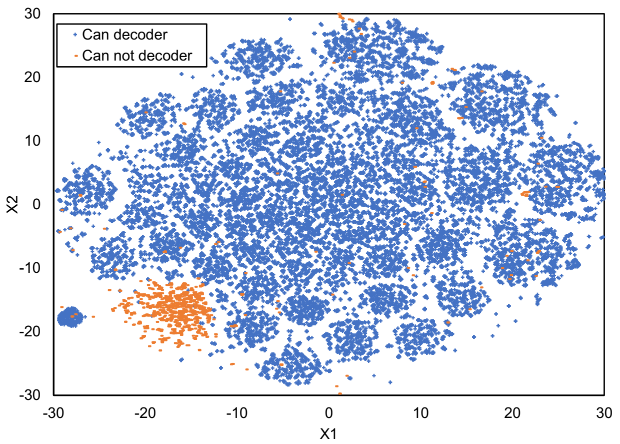

Fig. 7 Distribution of decodable and non-decodable materials. X1 and X2 are the two dimensions after dimension reduction.

DISCUSSION

The configurational phase space for new inorganic materials is immense. Forming four-component compounds from the first 103 elements of the periodic table results in more than 10^{12} combinations. Such a vast materials design space is intractable to high-throughput experiments or first-principle computations. On the other hand, current inorganic materials databases such as ICSD and Materials Projects all consist of only a tiny portion of the whole inorganic chemical space, which needs expansion for computational screening of new materials.

Here we proposed a GAN-based generative model for efficient sampling of the vast chemical design space of inorganic materials. Systematic experiments and validations show that our GAN models can achieve high uniqueness, validity, and diversity in terms of its generation capability. Our generative models can be used to explore the uncharted inorganic materials design space by expanding ICDS, material projects (MP), and OQMD databases. The derived expanded databases can then be used for high-throughput computational screening with higher efficiency than exhaustively screening billions of candidates[2]. While principles of charge neutrality and electronegativity balance[1] have been applied to filter out chemically implausible compositions for more effective search of new materials, such explicit composition rules are still too loose to ensure efficient sampling in the vast chemical design space for new materials search. Indeed, while the hypothetical materials with <5 elements can be enumerated (32 billion for 4-element materials with charge neutrality and balanced electronegativity), the design space of more elements can be challenging for which our GAN models can help a lot.

One limitation of our composition-based GAN model is due to the issue of possible polymorphism for a given composition: one composition may correspond to multiple structure phases and then multiple materials properties such as the bandgap. In this case, structural prediction and exploration is needed for our prediction compositions to search those phases to ensure a comprehensive inverse design process.

Our work can be extended in multiple ways. First, we found that MatGAN can learn chemical composition rules implicitly even though we did not explicitly enforce those rules into our GAN model. However, it is sometimes desirable to implement chemical rule filters to remove chemically invalid candidates, which can be easily implemented based on our matrix representation of materials. Another limitation in our current study is that we only considered the integer ratios of elements in compounds in our material representation while doped materials with fractional ratios are very common in functional materials such as lithium-ion battery material $LiZn_{0.01}Fe_{0.99}PO_4$, which is a doped cathode material. Our study can be extended by allowing real numbers on the representation matrix. However, considering the infinite possibility of doping ratios, our GAN method may need to work together with other sampling techniques such as genetic algorithms[30,31], genetic programming[32], and active machine learning for mixed parameter search[33,34] or the Bayesian optimization approach[33]. In addition, our current GAN models do not tell the crystal structures (lattice constants, space group, atomic coordinates, etc.) of the hypothesized materials. However, with sufficient computational resources, it is possible to exploit DFT-based computational software packages such as USPEX[35] or CALYPSO[36] to determine the crystal structure given a material composition and its stoichiometry. Our GAN models can also be used to work together with material structure generators[11].

METHODS

Datasets

We use a subset of inorganic materials deposited in the OQMD[15,17] database to train our AE and GAN models. OQMD is a widely used DFT database with crystal structures either calculated from high-throughput DFT or obtained from the ICSD[37] database. Currently it has 606,115 compounds. We use a similar screening criteria by Jha et al.[38] to choose the OQMD subset for GAN training: for a formula with multiple reported formation energies, we keep the lowest one to select the most stable compound. Single-element compounds are all removed along with materials whose formation energy is out of the range of $\{u-5\sigma, u+5\sigma\}$, where u and σ are the average and standard deviation of the formation energies of all samples in OQMD. Our final dataset, OQMD_L, has 291,884 compounds.

As comparison, we also train two GANs for the MP and ICSD databases, respectively. Both the MP and ICSD datasets here are prepared by removing all the single-atom compounds, the compounds which has any element with more than 8 atoms in their unit cell, and compounds containing Kr and He elements. The final MP dataset used here has 63,922 compounds. The final ICSD dataset used here has 28,137 compounds.

GAN neural networks training

We have optimized the hyper-parameters for training the GANs by setting the learning rate from 0.1 to 10^{-6} (each time decrease by 10-fold) and batch normalization size from 32 to 1024, and using different optimizers. We train our GANs using the screened samples from the OQMD, MP, and ICSD, and ICSD_filter database, which is an ICSD subset with all charge-neutral and electronegativity-balanced materials. These Wasserstein GANs are trained for 1000 epochs with the Adam optimizer with learning rate of 0.001 for the generator training and 0.01 for the discriminator training. The batch size for GAN training on OQMD is set to 512 while the batch sizes are set as 32 for GAN training on all other datasets. The AE is trained with the Adam optimization algorithm with a learning rate of 10^{-3} and batch size of 1024.

Training of bandgap prediction model

We choose 30,186 inorganic materials whose band gaps are >0 out of the 63,922 compounds in the selected MP dataset as the training samples to train the bandgap prediction model. Gradient-boosted decision tree (GBDT) machine learning model is then trained with the Magpie features. The learning rate is set as 0.06. The maximum tree depth is set as 20. The subsample is set to 0.4. The number of estimator is set to 100.

DATA AVAILABILITY

The training and test data are downloaded from the Materials Project database and OQMD database. Other results data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

Code and workflow developed in this study are available from the corresponding author upon reasonable request.

Received: 12 November 2019; Accepted: 4 June 2020;
Published online: 26 June 2020

REFERENCES

1. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem.* **1**, 617–627 (2016).
2. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
3. Jensen, W. B. & Jensen, W. B. The origin of the ionic-radius ratio rules. *J. Chem. Educ.* **87**, 587–588 (2010).
4. Ranganathan, S. & Inoue, A. An application of Pettifor structure maps for the identification of pseudo-binary quasicrystalline intermetallics. *Acta Materialia* **54**, 3647–3656 (2006).
5. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
6. Xue, D. et al. Advances and challenges in deep generative models for de novo molecule generation. *Wiley Interdiscip. Rev.: Computational Mol. Sci.* **9**, e1395 (2019).
7. Xu, Y. et al. Deep learning for molecular generation. *Future Med. Chem.* **11**, 567–597 (2019).
8. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design-a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
9. Ferguson, A. L. Machine learning and data science in soft materials engineering. *J. Phys.-Condens. Mat.* **30**, 043002 (2018).
10. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks, In *Proc. 34th Int. Confer. Mach. Learning*. Sydney, Australia **70**, 214–223 (2017).
11. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
12. Hoffmann, J. et al. Data-driven approach to encoding and decoding 3-D crystal structures. Preprint at https://arxiv.org/abs/1909.00949 (2019).
13. Nouira, A., Sokolovska, N. & Crivello, J.-C. Crystalgan: learning to discover crystallographic structures with generative adversarial networks. Preprint at https://arxiv.org/abs/1810.11203 (2018).
14. Sawada, Y., Morikawa, K. & Fujii, M. Study of deep generative models for inorganic chemical compositions. Preprint at https://arxiv.org/abs/1910.11499 (2019).
15. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Mater.* **1**, 15010 (2015).
16. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *Apl. Mater.* **1**, 1049 (2013).
17. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom* **65**, 1501–1509 (2013).
18. Doersch, C. Tutorial on variational autoencoders. Preprint at https://arxiv.org/abs/1606.05908 (2016).
19. Chen, X. et al. Infogan: Interpretable representation learning by information maximizinggenerative adversarial nets, in *Advances in Neural Information Processing Systems*, 2172–2180 (2016).
20. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein Generative Adversarial Networks, *Proc. 34th Int. Confer. Mach. Learning. PMLR* **70**, 214–223, (2017).
21. Pu, Y. et al. Variational autoencoder for deep learning of images, labels and captions. in *Advances in Neural Information Processing Systems,* 2352–2360 (2016).
22. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd Int. Confer. Int. Confer. .Mach. Learning.* **37**, 448–456 (2015).
23. Shamir, R. R., Duchin, Y., Kim, J., Sapiro, G. & Harel, N. Continuous dice coefficient: a method for evaluating probabilistic segmentations. Preprint at https://arxiv.org/abs/1906.11031 (2019).
24. Shamir, R. R., Duchin, Y., Kim, J., Sapiro, G. & Harel, N. Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations. Preprint at https://www.biorxiv.org/content/10.1101/306977v1 (2018).
25. Maaten, L. V. D. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
26. Ye, W., Chen, C., Wang, Z., Chu, I.-H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **9**, 3800 (2018).
27. Kang, S. & Cho, K. Conditional molecular design with deep generative models. *J. Chem. Inf. Model* **59**, 43–52 (2018).
28. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 16028 (2016).
29. Shao, S. et al. The exotically stoichiometric compounds in Al–S system under high pressure. *npj Computational Mater.* **6**, 1–6 (2020).
30. Atilgan, E. & Hu, J. First-principle-based computational doping of $SrTiO_3$ using combinatorial genetic algorithms. *Bull. Mater. Sci.* **41**, 1 (2018).
31. Atilgan, E. & Hu, J. A combinatorial genetic algorithm for computational doping based material design. In *Proc. Companion Pub. 2015 Annual Confer. Genetic Evol. Comp.* (ACM) 1349–1350 (2015).
32. Atilgan, E. *Computational Doping for Fuel Cell Material Design Based on Genetic Algorithms and Genetic Programming*. Ph.D. thesis, University of South Carolina (2016).
33. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
34. Lookman, T., Balachandran, P. V., Xue, D. Z., Hogden, J. & Theiler, J. Statistical inference and adaptive design for materials discovery. *Curr. Opin. Solid St. M.* **21**, 121–128 (2017).
35. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX—evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).
36. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: a method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
37. Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.* **23**, 66–69 (1983).
38. Jha, D. et al. Elemnet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 17593 (2018).

ACKNOWLEDGEMENTS

This work as partially supported by the National Science Foundation under grant numbers: 1940099, 1905775, OIA-1655740, and SC EPSCoR GEAR Grant 19-GC02 and by DOE under grant number DE-SC0020272. The views, perspective, and content neither necessarily represent the official views of the SC EPSCoR Program nor those of the NSF and DOE. The authors also acknowledge funding from the National Natural Science Foundation of China under grant number 51741101. This work is also partially supported by National Major Scientific and Technological Special Project of China under grant number 2018AAA0101803 and also by Guizhou Province Science & Technology Plan Talent Program under grant number [2017]5788. We would like to thank Zhuo Cao, Chengcheng Niu, and Rongzhi Dong for helpful discussions.

AUTHOR CONTRIBUTIONS

J.H. conceived the project. J.H. and Y.D. developed the methodology. Y.D. implemented the method. Y.D. and J.H. performed the calculations. Y.D., X.L., and Y.Z. prepared the figures. J.H., Y.D., X.L., S.L., and M.H. interpreted the results. J.H and Y.D. wrote the paper. J.H. and S.L. supervised the project. All authors reviewed and commented on the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00352-0.

Correspondence and requests for materials should be addressed to J.H.

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

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