THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

Active-Learning-Based Generative Design for the Discovery of Wide-Band-Gap Materials

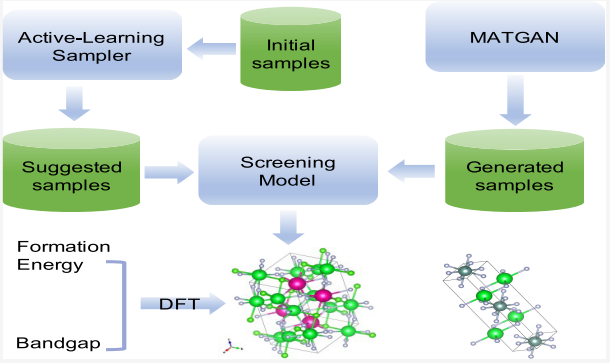
Rui Xin, Edirisuriya M. D. Siriwardane, Yuqi Song, Yong Zhao, Steph-Yves Louis, Alireza Nasiri, and Jianjun Hu*

Cite This: *J. Phys. Chem. C* 2021, 125, 16118–16128

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Active learning has been increasingly applied to screening functional materials from existing materials databases with desired properties. However, the number of known materials deposited in the popular materials databases such as ICSD and Materials Project is extremely limited and consists of just a tiny portion of the vast chemical design space. Herein, we present an active generative inverse design method that combines active learning with a deep autoencoder neural network and a generative adversarial deep neural network model to discover new materials with a target property in the whole chemical design space. The application of this method has allowed us to discover new thermodynamically stable materials with high band gap ($SrYF_5$) and semiconductors with specified band gap ranges ($SrClF_3$, $CaClF_5$, YCl_3, SrC_2F_3, AlSCl, As_2O_3), all of which are verified by the first-principles DFT calculations. Our experiments show that while active learning itself may sample chemically infeasible candidates, these samples help to train effective screening models for filtering out materials with desired properties from the hypothetical materials created by the generative model. The experiments show the effectiveness of our active generative inverse design approach. The source code can be freely downloaded from https://github.com/glard/Active-Generative-Design.

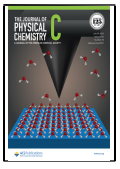


1. INTRODUCTION

Computational discovery of new materials with desired functions has big potential in a variety of industries. Currently, there are two major categories of approaches including (1) computational screening of existing materials in materials databases[1] to find candidates with new properties and (2) generation of new materials using rational design methods,[2] crystal structure prediction,[3] or generative machine learning (ML) models.[4,5] The computational screening approach, however, usually suffers from the limited known materials in terms of both quantity and diversity. This is especially true as most known materials in existing databases such as ICSD, Materials Project,[6] and OQMD[7] have only structural information annotated with a few density functional theories (DFT)[8]-calculated primitive properties such as formation energies or band gaps. However, a few of them are annotated with more interesting macroproperties such as thermal conductivity (<2000) and ion conductivity (<200), which makes it challenging to train a highly generalizable machine learning model for screening. The generation approaches also face a variety of challenges, such as the difficulty to generate diverse novel materials compositions and structures and the same issue of lacking high-quality property data for generated hypothetical materials. Essentially, both ML-assisted computational screening and generation suffer from very limited training data with scarce annotated property characterization data.

To improve the efficiency of materials screening, inverse design methods[9−16] have been proposed to search materials for a given target functionality, in which a sampling or optimization process is used to select a subset of the whole design space for locating the optimal design candidate. The sampling process in inverse design methods is usually guided by a global optimization algorithm[13,17] together with function performance evaluations via atomic simulation. Global optimization methods used in inverse design can be categorized into several types,[18] including gradient-based optimization methods; gradient-free optimization algorithms such as genetic algorithms, particle swarm optimization, and Covariance Matrix Adaptation Evolution Strategy (CMA-ES); neural network-based generative approaches; and Bayesian optimization methods using surrogate evaluation models. Depending on the discrete or continuous

Received: March 17, 2021
Revised: July 2, 2021
Published: July 20, 2021



https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

nature of the design targets, not all of these optimization methods can be applied directly. For example, the gradient methods cannot be easily applied to search the inorganic materials composition or crystal structure space as they tend to generate infeasible solutions during optimization processes.

Due to the strong physicochemical and geometric constraints among the atoms in periodic crystal materials, we find that gradient-based search operators or crossover/random mutation-based genetic search operators might still be too random in identifying the next feasible sampling point for expensive atomic simulation to obtain its property value. Instead, this exploration step can be guided by a machine learning model as the previously explored samples can be used to learn the structure–property relations. The resulting trained surrogate model can then be used to predict the properties of unexplored materials with a speed of 100 orders of magnitude faster than conventional DFT calculation methods. This has led to the emergence of active learning methods for materials design.[8,19,20]

The basic idea of active learning is its capability to explore the design space actively to search candidates with desired properties or identify informative samples for building property prediction models for labeling so that better prediction models can be trained with the existing and minimal number of newly acquired/labeled samples.[21] An active learning model is typically composed of the following components: an active learner with a specific machine learning prediction model and a query strategy to recommend new samples for labeling, and an oracle model to label the suggested samples.[22] The oracle model can be experimental or computational validations such as first-principles DFT calculation or even machine learning models. In general, active learning is very useful for tasks where data is not sufficient or expensive to acquire, which is the case for materials design problems. Since the active learner actively chooses the training data points, the number of data points needed for training a model based on active learning is typically lower or much lower than the number used by the passive learner. Active learning algorithms can be categorized by the choices of their components, especially the query strategies, which can be classified into uncertainty-based, ensemble-based, and diversity-based approaches. The key idea of an uncertainty-based method is to query those areas about which predicted value the active sampler is least certain. Uncertainty sampling strategies can be employed not only with probabilistic models, which help sample the instance whose best labeling is the least confident, but also nonprobabilistic models, which can be modified to have probabilistic output. For example, nearest-neighbor-based classifiers in which neighbors are voting on class labels, and the ratio of the vote on a targeted class represents the posterior probability.[23] Moreover, a semisupervised-learning-based active learning algorithm uses variational autoencoder (VAE) and adversarial network to implicitly learn the sampling mechanism by utilizing the representation outliers and uncertainty together.[24] Ensemble-based methods maintain a series of machine learning models and help sampler decide valuable query points with which these models most disagree. ACTIVE-DECORATE,[25] which uses DECORATE committees to select good training examples, outperforms traditional ensemble methods such as Bagging and Boosting. DECORATE algorithm is designed to use additional artificially generated training data from data distribution and label them different from current ensemble predictions to generate highly diverse ensembles. Among all of these methods, uncertainty-based active learning has been most successfully applied to materials design.

Despite the success stories of a few active-learning-based materials design cases, there exist several limitations to improve. For example, in ref 8, active learning has been applied to find high-band-gap materials. Their model is initially trained with only around 2% of the entire search space, and 20 new samples suggested from the optimization schemes, are added at each optimization cycle. Their experiments show that the active learning approach is able to identify the materials with the highest property value after searching only around 7.0 and 7.7% of the entire database. However, as a demonstration paper, this work does not identify any new materials and have very limited design space compared to the real-world active inverse design problem. In addition, current active learning for materials design studies have been dominated by screening existing materials databases such as ICSD or Materials Project,[8] which limits their search space to a small portion of the vast chemical design space, which may unexpectedly exclude more promising candidates.[26]

We propose a method to combine active learning with generative machine learning to build the active generative inverse design framework for new materials discovery and apply it to the discovery of materials with a desired band gap. Compared to the existing active learning search of high-band-gap materials, which is focused on exploring the known database, our study focuses on an active-learning-based exploration of the unknown design space for high-band-gap materials. The framework includes four components, including a materials composition generator, an autoencoder sampler, a band gap predictor, a Bayesian optimization-based sampler, and a substitution-based structure predictor. Our composition-based generative machine learning model[5] can generate millions of new compositions that satisfy basic chemical rules such as charge neutrality and balanced electronegativity and tend to have low free energy. We then build a composition-based band gap prediction model using a Random Forest regressor. A Gaussian process regressor-based Bayesian optimization algorithm is used to train an ML-based sampling model, which suggests the next sampling points in the active learning/search process to maximize expected property improvement. To verify the predicted materials, we use substitution-based structure predictors to get candidate structures, which can then be used to verify their structural stability and their band gaps. Our active generative design (AGD) method has allowed us to find a high-band-gap material $SrYF_5$ with a band gap of 6.42 eV and seven semiconductor materials with desired band gap values.

Our contributions include the following:

- We propose a novel active generative inverse design approach for discovering new materials in the whole design space rather than in a given database. We find that samples suggested by the active learning model may not be chemically valid. But they are useful for building good property prediction models.
- We show that active learning can be used to increase the property prediction model's capability to screen out high-band-gap materials by sampling informative data points for property evaluation, which is highly desirable for screening materials with small datasets of property annotations.
- Our evaluation experiments show that active-learning-based sampling can be combined with generative models to speed up new materials discovery. Here, we train an ML

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

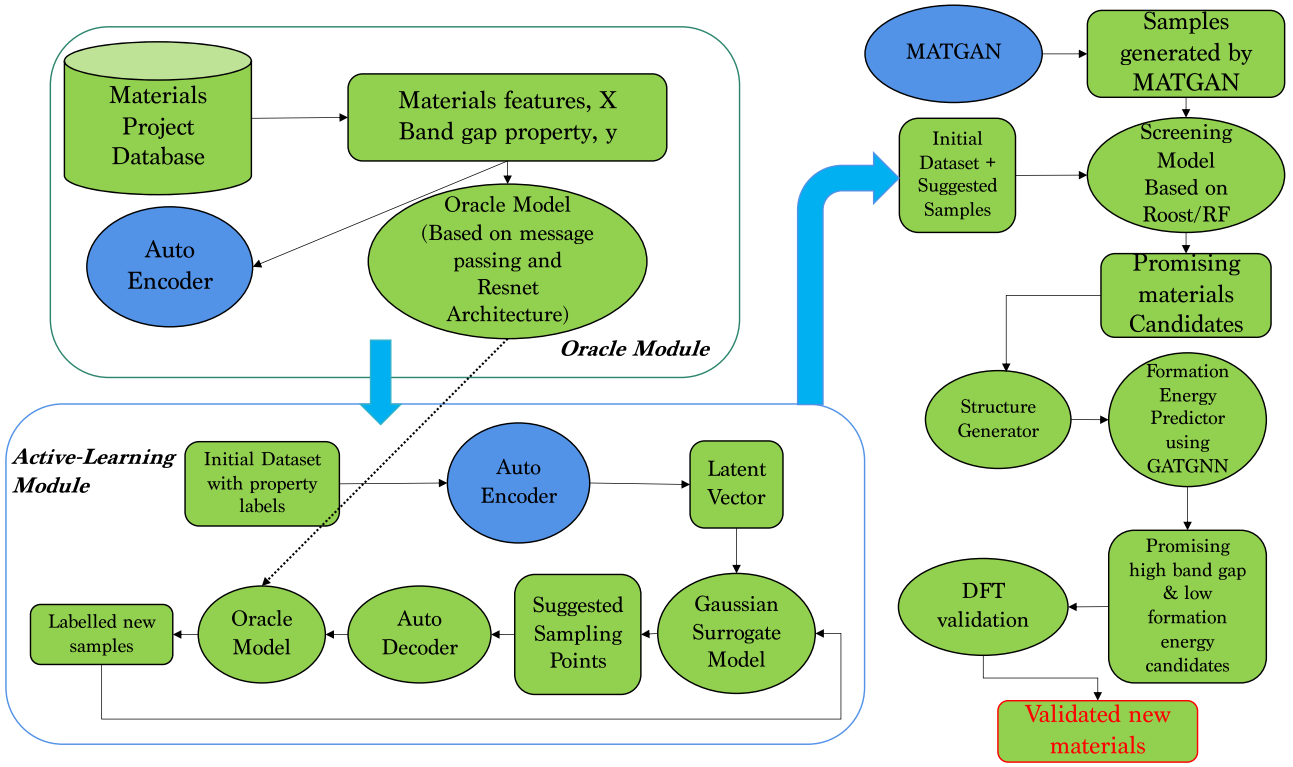


Figure 1. Framework for active generative design for materials discovery. The Oracle model is trained with all dataset for the purpose of model evaluation. The active learning module is used to recommend sampling points for expensive property calculation, which are then used for building the ML screening model. The MATGAN is used to generate candidate materials, which are then filtered by the ML screening model.

model from samples traversed by the active learning search process for screening hypothetical material candidates created by the generative models.

- Our new pipeline has been able to discover 11 semiconductor materials (Table 2) and two high-band-gap materials ($SrYF_5$, $RbSr_2ClF_4$) with DFT-validated band gaps of 6.42 and 5.64 eV, respectively.

2. MATERIALS AND METHODS

2.1. Active Generative Design (AGD) Framework for Materials Discovery. The main idea of our AGD framework is to use the active learning approach to learn a highly accurate property prediction model and combine it with a generative model (e.g., an autoencoder model or a Generative Adversarial Network (GAN model)). Our whole active generative design framework is shown in Figure 1. It is composed of three modules: an active learning model for training a high-performance property predictor using limited sample labeling, and a candidate generation model, and a screening model for the discovery of new materials. For the purpose of fast benchmark studies here, a machine-learning-based oracle model is trained to simulate the DFT calculations of band gaps in this study.

The active learning module is the part where data augmentation via active exploration of the design space is performed to acquire labels of the most informative sampling points to improve the performance of the machine-learning-based Screening Model. The generative module uses VAE or GAN models to generate candidate materials and feed them to the Screening model trained by the enhanced dataset to filter out a number of promising materials candidate formulas. Their corresponding candidate structures will subsequently be generated by exploiting their similarity with existing structures using the element substitution approach. Formation energy will also be predicted using Global Attention Graph Neural Network (GATGNN).[27]

Within our active generative design framework, there are two ways to discover hypothetical materials:

- (1) ALS: direct active learning search in the whole chemical space, which is done by combining the active learning sampling model with the autoencoder decoding model.
- (2) AL+RF/Roost with MATGAN: active-learning-based screening of candidates generated by the MATGAN generative model. Active learning is used to create labeled samples for training ML screening models using RF or Roost, which are then used to filter the samples generated by MATGAN.

2.1.1. Oracle Module. The Oracle Module is a machine learning model trained with material features and band gap labels of the whole dataset in our experiments. This module aims to simulate the DFT function and facilitate the prediction of band gap property based on a vector of material features. In practice, this Oracle model should be replaced with DFT calculations or experiments. The descriptors for our Oracle model are calculated using Roost (Representation Learning from Stoichiometry).[28] Roost is a machine learning algorithm that learns the stoichiometry-to-descriptor map directly from data, the key idea of which is each material's composition is represented as a dense weighted graph. Roost uses Matscholar embedding[29] for embedding and CGCNN[30] for internal representation. The output network used for the reference model is a Resnet with five hidden layers and ReLU activation functions.

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

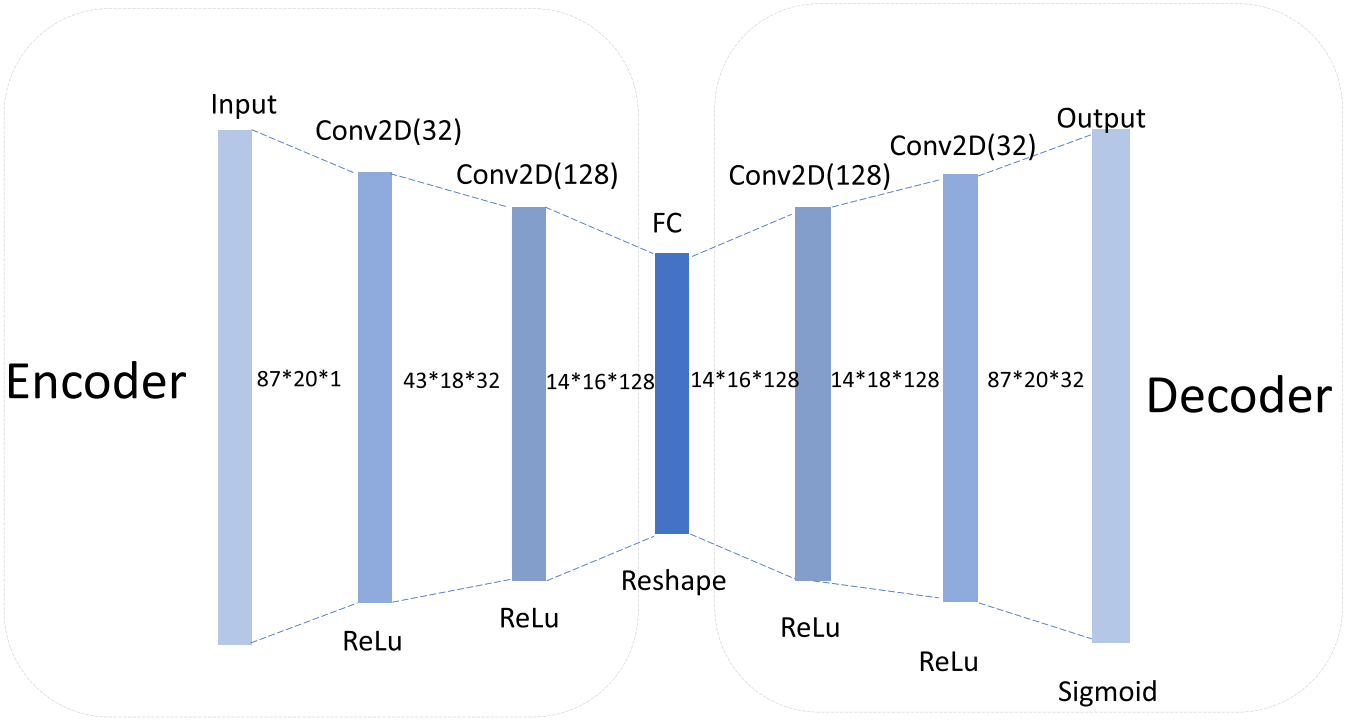


Figure 2. Autoencoder model.

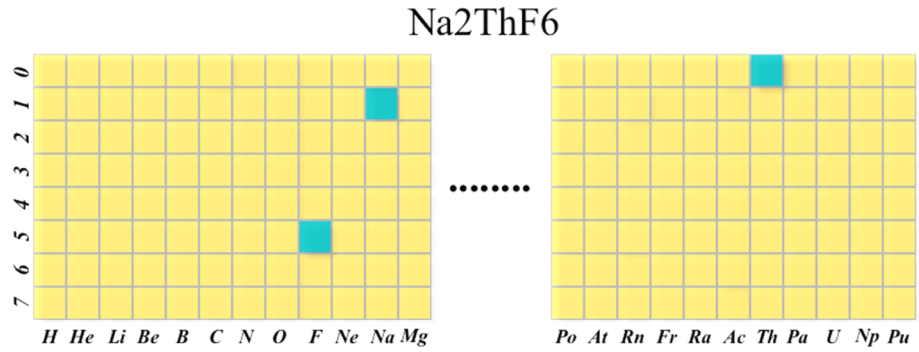


Figure 3. One-hot encoding of material compositions.

2.1.2. Active Learning Module. The function of the Active learning Module is to actively explore the whole design space to identify the informative sampling points and obtain their property annotations, which are fed to train a highly accurate property prediction model. In previous active-learning-based design studies such as ref 8, the samples are directly represented as 145 common chemical features, which are used to train the Gaussian Surrogate model. New sampling points are not generated but selected from existing entries in a given database. However, such physical feature representation of materials is not applicable for active learning in the uncharted design space because the sampled feature representations cannot be easily mapped back to valid materials formulas/structures. Actually, this reverse mapping procedure usually makes it problematic for the Gaussian Surrogate model to generate meaningful samples that can be decoded into chemical formulas. To address this representation issue, we introduce the AutoEncoder deep neural network model as a latent vector generator. It can be used to convert a chemical formula into a 128-dimension latent vector and vice versa. Thus, formulas are not only well represented but also provide latent vector boundary for each dimension which makes it easier for the optimization process. To get the property annotation by the Oracle model, the recommended sampling points from the Surrogate model will be first decoded into a formula, and then its Magpie descriptor features are calculated, and the band gap property will be predicted by the Oracle model.

Latent vectors as generated above, along with their property annotations, will be fed as input to the Gaussian Surrogate model, which will give uncertainty-based sampling point suggestions for Oracle Module to evaluate. This suggested vector has to be converted back to a chemical formula to get material features. However, cases might occur when some suggested latent vectors cannot be decoded correctly into formulas and will negatively affect the design space exploration performance of the active learning module. Afterward, labeled new samples will be sent back to the Gaussian Surrogate model for a new round of recommendations. It should be noted that different initial datasets and the Gaussian Surrogate model constructed both have an influence on the sampling points, which affect the final performance of the ML screening model trained with these data points.

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

2.2. Autoencoder Neural Network for Latent Space Sampling. Standard autoencoder is a popular representation learning and dimension reduction method, which is composed of two parts: an encoder and a decoder. Our autoencoder model is used to map material compositions into a continuous space, which can be used by the Bayesian optimizer to sample the design space. The sampled points can then be decoded back into materials compositions. In our model, the encoder has two convolutional layers followed by a fully connected layer and a decoder consists of a fully connected layer and three convolution layers. The autoencoder (AE) model is shown in Figure 2. The ReLU is used as activation function in each neural layer except the last layer of the decoder for which the Sigmoid function is used instead. We use the Adam optimizer to train the model parameters and the learning rate is set as 0.001. We use the cross-entropy loss for the training. The batch size is set as 128. The autoencoder is trained with 66 324 inorganic materials selected from the Materials Project database.

The material compositions are represented using the one-hot encoding. Through the simple statistical calculation of the materials in the MP dataset, 87 elements are found and each element usually has less than 20 atoms in any specific compound/formula. We then represent each material as a sparse matrix with 0/1 cell values. Each column represents one of the 87 elements, while the column vector is a one-hot encoding of the number of atoms of that specific element. The whole encoding scheme is shown in Figure 3.

2.3. Active Learning Module for Efficient Sampling. The active learning module in our pipeline is used for efficient exploration of the design space to identify informative samples and obtain their property annotations, which allows generating more training samples to train an accurate property prediction model. The sampled materials can also be used for screening candidates from them with desirable properties. We use Bayesian optimization (BO)[31] as the active learning sampling strategy to explore the huge design space.

BO is an efficient global optimization (EGO) algorithm aiming to find the maximum value of a costly unknown function in as few iterations as possible. BO builds Gaussian process regression, a surrogate model, to not only evaluate the objective function but also quantify the uncertainty with each prediction. Acquisition functions such as expected improvement (EI)[32] and upper confidence bound (UCB)[33] are conducted to find the next point to evaluate.

$$\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \quad (1)$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the mean and standard deviation of GP posterior at point x, respectively, and Φ and ϕ are the cumulative distribution function (CDF) and probability density function (PDF) of the standard normal distribution, respectively. Z is denoted in eq 2. $f(\mathbf{x}^+)$ represents the best observed value, whereas x is the corresponding location.

The first term in eq 1 denotes exploitation, and the second summation term is the exploration term. EI can adjust the level of balance between exploitation and exploration through changing parameter ξ. The higher the ξ is, the lower the level of importance of posterior mean value improvement, while the algorithm will be more inclined to explore unknown areas.

$$Z = \begin{cases} \dfrac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \quad (2)$$

Our framework adopted the acquisition function UCB in eq 3 to probe promising points. Intuitively, UCB explores the design space to locate the max value by considering both expected performance and uncertainty across all solutions. The κ factor here enables fine adjustments between exploration and exploitation.

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}} (\mu_t(\mathbf{x}) + \kappa\sigma_t(\mathbf{x})) \quad (3)$$

Fernando Nogueira's package[34] is used as the sampler to provide suggestions about which latent vectors will be generated and probed during the search process. Depending on the experiments, different numbers of initial sampling points (prior) are used to build the Gaussian process surrogate model. In every iteration, the sampler will suggest 20 sampling points. Only those candidates that can be decoded into a valid material composition will be registered and used to update the Gaussian process model. The sampler will stop sampling when the set budget is reached.

2.4. Template-Based Structure Prediction and DFT-Based Validation. To validate the band gap property of promising candidates generated by our framework, we perform a template-based structure prediction over the generated compositions to create their structures for DFT calculation after charge neutrality and electronegativity integrity checking.

We first measure the similarity between a candidate chemical formula and all of the formulas from the Materials Project (MP) database using Earth Mover's Distance (EMD).[35] EMD computes the ratio of each of the elements and the absolute distance between the elements on the modified Pettifor scale between two compositions, which leads to greater alignment with chemical understanding than the Euclidean distance. The MP formulas are then ranked by their EMD distance, and the CIF file of the most similar formula is identified as the structure template for element substitution to generate the structures for the candidate compositions. The CIF files of these new structures will then be used as input for a pre-trained Global Attention Graph Neural Network (GTGNN)[27] to predict their formation energy. Only candidates with negative formation energy will be kept. Finally, DFT calculations will be performed over these filtered structures to calculate their band gaps and formation energy. Phonon calculation is also applied to one of the discovered materials to check its thermodynamic stability.

3. EXPERIMENTS

3.1. Band Gap Dataset. We download the Materials Project dataset from http://materialsproject.org and remove duplicate composition entries to get 42 667 samples with band gap data. We call this the MP dataset. The distribution of band gaps of the MP dataset is shown in Figure 4.

3.2. Experimental Design. To validate the performance of our active generative design framework, 1 180 000 potential material compositions are generated via the MATGAN network.[5] Bayesian optimization with the UCB utility function is used as active learning to explore promising sampling points for machine learning model training. κ is set to 100 since the goal is to generate diverse samples. The Roost band gap prediction model is trained based on the original and extra generated

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

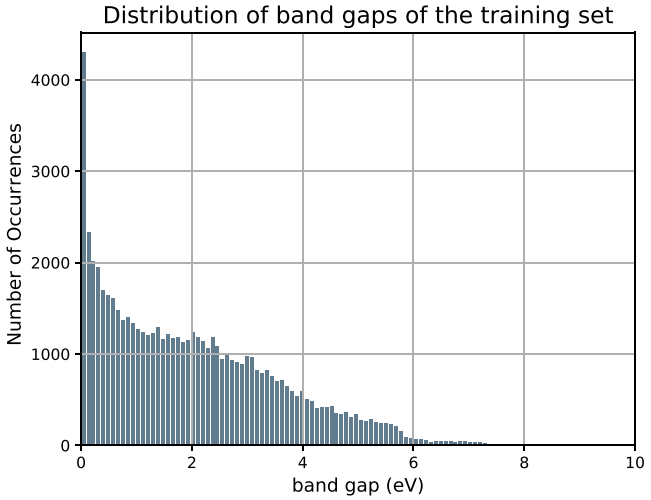


Figure 4. Distribution of band gaps of the training set.

sampling points so as to screen high-band-gap materials out of these potential materials. To better evaluate the effectiveness of our AGD framework, we designed four experiments according to the availability of the different amounts of training data. Each model is evaluated by repeating the experiment 10 times to report statistically significant results. In each experiment, we split the whole dataset into 60% for training, 20% for validation, and 20% for testing, respectively.

(1) Exp1-oracle model: 42 667 materials from the Materials Project database are used for training the Oracle model for band gap prediction, which is then used to screen high-potential-band-gap materials from MATGAN-generated candidates. This is the ideal case where there are abundant annotated materials samples for training an ML property prediction model for screening.
(2) Exp2: We randomly select 1000 labeled samples from the MP dataset to train a baseline ML model (Exp2 BS). We then use the Gaussian Surrogate model for active learning. After 1000 sampling points are suggested by the Bayesian optimizer, 741 samples can be decoded into valid compositions. We use these newly explored samples together with the 1000 initial samples to update our training dataset and train an updated Screening Model (Exp2 AL). Both band gap prediction models will be evaluated and used for screening MATGAN-generated candidates.
(3) Exp3: Only 300 samples are randomly selected from the MP dataset for the baseline ML model (Exp3 BS) training and Gaussian Process initialization. We then generate 1000 new samples using the active learning model, which leads to 853 valid compositions. In total, 1153 samples are used to train the updated ML model (Exp3 AL). We compare the performance between the baseline screening model (BS) and the auxiliary Active Learning trained screening model (AL) as we do in Exp2.
(4) Exp4: Only 300 samples whose band gap values are below 3.0 eV are selected for active learning. This control experiment aims at validating the performance of our AGD framework when only a small, poor-quality dataset is available.

To compare the performance of the model performances trained with a different number of samples, as shown above, we pick the 8535 test samples from the Exp1 and remove the entries that appear in the Exp2 and Exp3 experiment datasets. We then select 500 samples from the remaining samples as the test set for evaluating all of the models.

Additionally, a control experiment is conducted in which we adjust the parameter κ of the utility function UCB to inspect the active learning behavior between exploration and exploitation.

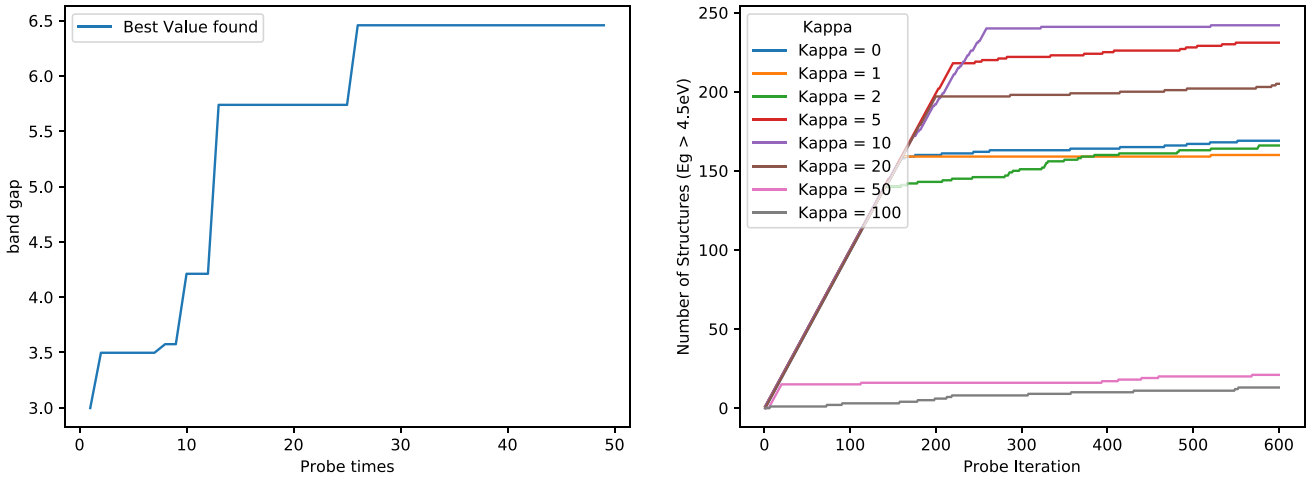
3.3. Results. *3.3.1. Prediction Model Capability with Active Learning.* Table 1 shows model performance and the number of candidate materials with a band gap greater than 6.0 eV found with or without active learning. Exp1 has the most available training data, which results in an average R^2 performance of 0.735 when evaluated on its 20% hold-out test set. The baseline models trained with 1000 and 300 samples in Exp2 and Exp3 have corresponding R^2 scores as low as 0.361 and −0.243, showing the under-performance due to lack of sufficient training samples. The negative R^2 score of the baseline model performance of Exp3 BS shows that 300 initial samples are not sufficient to train a reliable model. In contrast, 1000 initial samples can train a much better model. However, after we apply the active learning process, the great performance gain is achieved as the model metric R^2 scores are improved from 0.361 to 0.611 in Exp2 and from −0.243 to 0.757 in Exp3 when evaluated using the random hold-out test over the dataset composed of initial samples and newly sampled points. However, the improvements are a little over-estimated here. We can observe that the performance improvement when evaluated with the external 500-sample test set. The improvement due to active learning is reduced to 0.19 for Exp3 AL and is only 0.005 for Exp2 AL. After close examination, we find that this is due to the mismatch of the training set distribution and the external test set distribution: most of the samples suggested by active learning during the search process have high band gap values greater than 2 eV while the majority of the whole dataset from which the 500-sample test set are drawn from have band gap values between 0 and 2 eV (see Figure 4). This shows that the samples generated during the active-learning search are highly biased toward candidates with higher band gaps.

During active learning of Exp2 and Exp3, we find that 853 decodable new materials out of 1000 Bayesian optimization iterations are found and evaluated in Exp3 AL, while 741 new materials are found in Exp2 AL. The Exp2 AL finds fewer new materials given the same budget because more initial samples will lead to a better Gaussian process model with less uncertainty

Table 1. Average Model Performance and Screening Results

exp no.	no. of samples	R^2 (hold-out)	R^2 (test set)	MAE	RMSE	no. of candidates	recovery rate (%)
Exp1	42 677	**0.735 ± 0.021**	**0.850 ± 0.023**	**0.505 ± 0.016**	**0.826 ± 0.034**	102	100
Exp2 BS	1000	0.361 ± 0.022	0.409 ± 0.016	0.882 ± 0.024	1.208 ± 0.021	123	51.96
Exp2 AL	1741	**0.611 ± 0.009**	0.414 ± 0.013	**0.539 ± 0.008**	**0.869 ± 0.010**	**171**	**78.43**
Exp3 BS	300	−0.243 ± 0.082	0.227 ± 0.042	1.202 ± 0.057	1.633 ± 0.054	33	2.94
Exp3 AL	1153	**0.757 ± 0.015**	**0.417 ± 0.005**	**0.309 ± 0.019**	**0.592 ± 0.019**	**81**	**57.84**

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128



(a) Band gap trace of the sampled materials during the search process (b) Exploration vs Exploitation (UCB utility function).

Figure 5. Active inverse design process.

so that the optimizer tends to suggest smaller steps on the latent vector. Such latent vectors with small changes tend to result in decoding into the same materials.

In Exp4 AL, we perform active learning based on randomly sampled 300 materials whose band gap is below 3.0 eV. Figure 5a shows our framework can easily find high-band-gap potential materials (>6.0 eV) within 30 iterations. However, these found materials may not be valid if chemical rules are applied. As a result, most of these generated labeled samples will function as data augmentation measurements for screening model training.

3.3.2. Exploration versus Exploitation in Active Learning: Control Experiments. We conduct several control experiments to find the relationship between exploration and exploitation during the active learning process. To find the suitable trade-off between them, we conduct experiments using the Upper Boundary Confidence (UCB) acquisition function. The κ parameter can be used to control the exploration rate. A larger κ leads to more exploration, while a smaller one leads to more exploitation.

The relationship between exploration and exploitation is shown in Figure 5a. The blue line represents pure exploitation, which flattens after about 150 iterations since nearly all candidates with high expected values have been utilized without exploration. As κ gradually increases, which means UCB explores more, the algorithm will more likely jump out of local minimum. As a result, the number of identified high-band-gap structures almost always increases when we increase κ from 0 to 10. However, the number of found structures decreases when κ increases from 10 to 100 as active learning continues to lean towards exploration. The gray line (κ = 100) and pink line (κ = 50) become flatten almost from the beginning since uncertainty has a rather large enhancement as shown in formula 3. As a result, active learning arrives at its trade-off and maximizes the discovery of promising structures when we set κ to 10.

3.3.3. Finding New Materials with Targeted Band Gaps Using the Active Generative Design Framework. To understand the effect of using active learning, we apply the trained ML models to screen the MATGAN-generated hypothetical materials and inspect their true band gaps calculated using the Oracle model.

The last two columns of Table 1 show the number of candidate compositions with predicted band gaps greater than 6 eV and the recovery rate measured by the overlap percentage of candidates with respect to the candidates screened out by the Oracle model (Exp1). The recovery rate column of Table 1 also shows models trained with active learning (AL) augmented data achieve a much higher recovery rate than the baseline model (BS). Specifically, with a 1000 search budget, Exp2 AL and Exp3 AL improved the recovery rates of the baseline models by 26.47 and 54.90%, respectively, indicating that active learning can improve the screening capability of high-band-gap materials with the ML models trained with suggested samples and initial samples, although their R^2 score, as evaluated by the external dataset, is improved just a little.

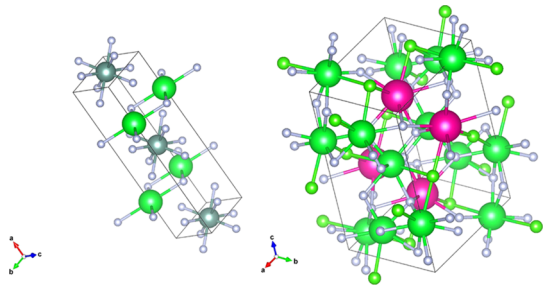
From the screening sets, we have found a high-band-gap material $SrYF_5$ in Exp1, Exp2 AL, and Exp3 AL whose predicted band gaps are 6.71, 7.02, and 7.11 eV, respectively. We use the element substitution to get its structure (Figure 7a) and calculate its band gap using DFT calculation, which shows a band gap value of 6.42 eV. However, $SrYF_5$ does not appear in the screening results of models from Exp2 BS and Exp3 BS, which are the baseline models constructed without conducting active learning. This demonstrates the benefit of active learning in new materials discovery.

Additionally, our active learning process has found a new material $RbSr_2ClF_4$ (structure shown in Figure 7b) directly during the active learning process. Its DFT-validated band gap value is 5.64 eV. This material is a minor variation of the material $RbSr_2Cl_5$ (band gap: 5.20 eV), a member of the initial dataset for active learning training. Altogether, the results show that our active generative design framework is capable of working well when only a small training dataset is available by efficiently exploring the design space and improving ML model performance.

3.3.4. Finding New Materials with MATGAN+Screening Approach. The main idea of active generative design is to use active learning to train a good screening property prediction model, starting with limited training samples. When there is a sufficient number of training samples, one can simply build an ML model as the filter to screen the hypothetical samples

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

generated by GAN models. When there are limited labeled training samples, it is better to use active learning to improve the ML model using as few expensive (e.g., DFT) sample annotation as possible. To prove the utility of active learning for such screening-based new materials discovery, we applied it to the high-band-gap materials screening problem starting with only 300/1000 labeled samples (Figure 6).



(a) Structure of $SrYF_5$ (b) Structure of $RbSr_2ClF_4$

Figure 6. Structures of the discovered new materials.

We first use the MATGAN, a generative adversarial network (GAN) for efficient generation of new hypothetical inorganic materials.[5] Trained with materials from the ICSD database, the MATGAN model generates 2.99 million inorganic formulas, most of which are charge-neutral and electronegativity-balanced. To simplify the screening process, we trim the generated formula set by filtering out those formulas with the total atom numbers more than 20 or the number of elements greater than five, or the number of atoms for each element more than seven. After checking chemical validity, the total number of candidates for screening is 1.18 million, and we call these samples the MATGAN dataset. We then use the whole MP dataset to train a Random Forest model to screen this dataset for potential semiconductors with a band gap between 2.0 and 4.0 eV. Similarly, we use the initial samples in Exp1, and Exp2 plus the active learning explored samples to train two Roost ML models and use them to screen the MATGAN dataset for band gaps greater than 6.0 eV. From this final candidate list, we select 18 for DFT formation energy and band gap validation.

3.4. DFT-Based Calculations. The density functional theory (DFT)-based first-principles calculations were carried out based on the Vienna Ab initio Simulation Package (VASP).[36−39] The projected augmented wave (PAW) pseudopotentials were considered for the electron−ion interactions with 400 eV plane-wave cutoff energy.[40,41] The exchange−correlation functional was treated using the generalized gradient approximation (GGA) based on the Perdew−Burke−Ernzerhof (PBE) method.[42,43] The energy convergence criterion was set as 10^{-5} eV. The atomic positions were optimized until the forces become less than 10^{-2} eV/Å. The Brillouin zone integration for the unit cells was performed using the Γ-centered Monkhorst−Pack *k*-meshes. The Formation energies (in eV/atom) of several materials were calculated based on eq 4, where E[material] is the total energy per unit formula of corresponding material, $E[A_i]$ indicates the energy of the ith element of the material, x_i represents the number of A_i atoms in a unit formula, and N is the total number of atoms in a unit formula ($N = \sum_i x_i$). Those values are reported in Table 2. $CaClF_5$ (−2.09 eV/atom), $SrYF_5$ (−4.10 eV/atom), and $SrClF_3$ (−2.55 eV/atom) show very low formation energies, implying our model is capable of finding new

Table 2. DFT-Verified Formation Energies and the Band Gaps of Several Semiconductors

material	E_{form} (eV/atom)	E_{gap} (eV)	method
$SrYF_5$	−4.10	6.42	AL+Roost with MATGAN
$RbSr_2ClF_4$	−3.35	5.64	Direct Active Learning Search
SiO_2F	−1.69	3.99	Direct Active Learning Search
PO_2F	−2.10	3.84	Direct Active Learning Search
SH_2O_3	−0.10	3.74	Direct Active Learning Search
SiF_3	−4.27	3.79	Direct Active Learning Search
NaF_5C	−1.78	4.80	AL+RF with MATGAN
$CaClF_5$	−2.09	3.46	AL+RF with MATGAN
YCl_3	−1.94	2.18	AL+RF with MATGAN
$SrClF_3$	−2.55	3.18	AL+RF with MATGAN
$SrCl_2F_3$	−1.91	2.27	AL+RF with MATGAN
AlSeCl	−0.58	3.70	AL+RF with MATGAN
As_2SO_3	−1.16	2.72	AL+RF with MATGAN

semiconductor materials, which are thermodynamically highly stable relative to the parent compounds of their elements.

$$E_{\text{form}} = \frac{1}{N}\left(E[\text{material}] - \sum_i x_i E[A_i]\right) \quad (4)$$

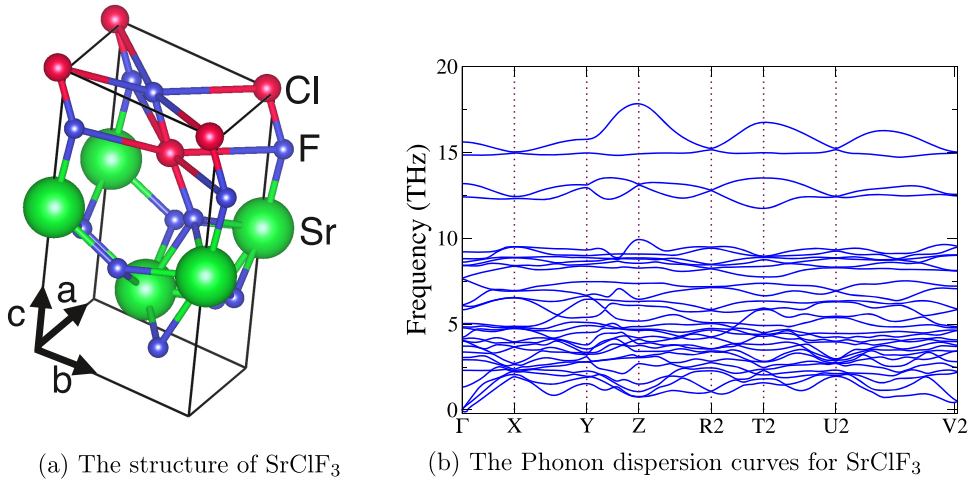
Moreover, we find that $SrClF_3$ is a stable semiconductor with $P1$ space-group symmetry (Figure 7). The unit cell lengths of the material are $a = 4.44$ Å, $b = 4.75$ Å, and $c = 8.30$ Å, while the unit cell angles are $\alpha = 102.3$, $\beta = 90.1$, and $\gamma = 90.1$. The positive frequencies in the phonon calculations (see Figure 7b) show that the material is dynamically stable at 0 K temperature. The energy above hull, which is 0.046 eV/atom, is obtained against the competing phases SrF_2 and ClF using Pymatgen code.[44]

We also carry out density functional perturbation theory (DFPT) calculations as implemented in VASP code for studying the mechanical properties of $SrClF_3$. Due to the $P1$ space-group symmetry of this material, there are 21 independent elastic constants. The calculated $C_{11} = 64.19$ GPa and $C_{12} = 62.87$ GPa elastic constants are almost equal because of $a \approx b$, and $\alpha = \beta$ lattice parameter relationships. Bulk modulus (41.26 GPa), Shear modulus (15.61 GPa), and Young's modulus (41.587) were calculated based on the Hill approach.[45] When an infinitesimal strain (ϵ) is applied on a material, the total energy of that material can be written as $E = E_0 + (1/2)V_0\sum_{i,j=1}^{6} C_{i,j}\epsilon_i\epsilon_j + O(\epsilon^3)$. Here, C represents the matrix of second-order elastic constants. The Born elastic stability conditions show that the matrix should be definite positive, all eigenvalues of C should be positive, and all of the principal components should be positive. Even though most crystal symmetries have simple relationships between the elastic constants to show whether the materials are mechanically stable, triclinic systems have very complicated equations because of having 21 independent elastic constants.[46] Thus, we used VASPKIT[47] code to analyze the elastic properties, and it confirms that $SrClF_3$ is very likely to be a mechanically stable semiconductor.

4. DISCUSSION

Our active learning experiments have shown that it can find new materials with high band gap values as validated by DFT, demonstrating that active-learning-based generative design has the capability to discover new material with target properties using guided/directed search. Figure 8 shows how active learning explores the inorganic material space. Active-learning-probed samples (circles) are obviously more widely distributed

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128



(a) The structure of $SrClF_3$ (b) The Phonon dispersion curves for $SrClF_3$

Figure 7. Structure and phonon dispersion of $SrClF_3$.

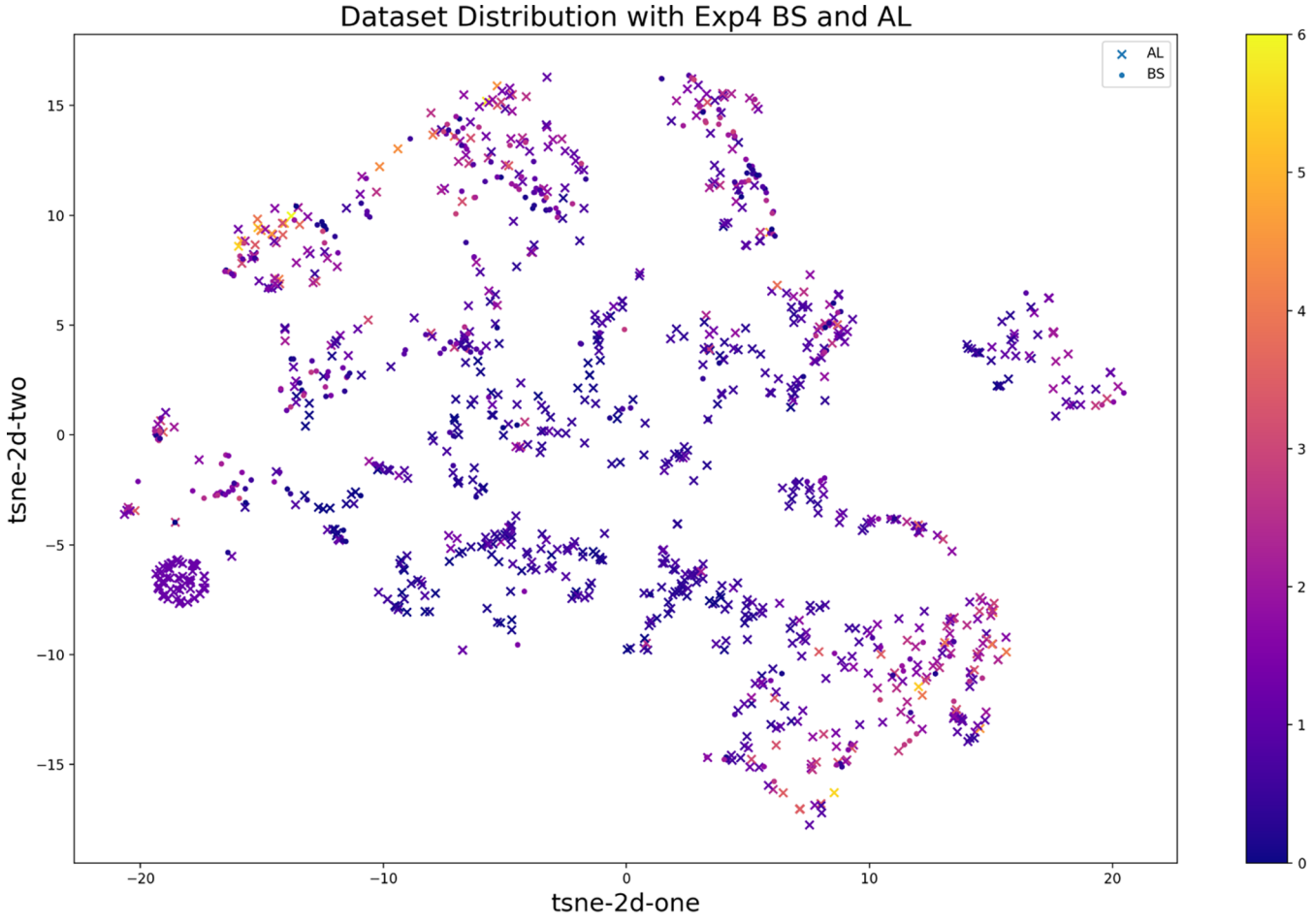


Figure 8. Comparison of the sample distributions of the baseline training set and the AL-enhanced dataset using t-SNE visualization. It is found that the samples traversed by the active learning algorithm are much more diverse and span the majority of the design space.

than the original training set (+). However, there are several aspects to improve for our active generative design framework.

First, currently, we are using Bayesian optimization as the efficient global optimization (EGO) strategy to search the 128-dimension design space containing the entire inorganic materials. Yet, many latent representations sampled during active learning search cannot be decoded into valid chemical formulas, which has caused a significant waste of the computational budget and decreases the efficiency of the Bayesian optimization sampler. Additional physicochemical constraints need to be incorporated into the sampling strategy so as to achieve more efficient sampling. Second, we encountered the out-of-distribution prediction issue of standard machine learning: our random forest/Roost models cannot predict band gap values that are larger than the maximum band gaps in the training set.[48,49] Surrogate ML models with higher

https://doi.org/10.1021/acs.jpcc.1c02438
J. Phys. Chem. C 2021, 125, 16118−16128

generalization capability are needed to future improve the prediction accuracy of our screening ML models.

Another key observation of our active-learning-based screening experiments is that the machine learning models such as the graph neural network model Roost model do not perform too well on the whole dataset since the dataset is highly biased (too many data points around 0 eV). However, the ML screening model trained by the augmented samples is capable of identifying high/wide-band-gap materials during the screening process. The reason is that the suggested samples by the active learning tend to focus on high-band-gap materials.

We also tried to directly use active learning to search the whole design space for discovering high-band-gap materials, which indeed has found several new materials such as $RbSr_2ClF_4$ and four others. Compared to previous studies which focus on screening known materials databases, this approach has the advantage of search in the complete chemical space. However, we find that many candidates generated via this approach by the active learning model do not satisfy basic physical requirements such as neutrality and balanced electric negativity. However, these suggested samples are useful for building an ML-based screening model as they have effectively explored the design space. The resulting model can then be applied to the generated samples by MATGAN. Therefore, for semiconductor materials design, we use the MATGAN+screening approach, which is more suitable to discover semiconductors since the intuition of optimization is to find the maximum or minimum of a black-box function.

5. CONCLUSIONS

We propose an active-learning-based generative design approach for the discovery of new inorganic materials and apply it to the discovery of high-band-gap materials and semiconductor materials. Our active-learning-based screening is unique due to its search in the whole design space rather than known databases as done in the previous studies. Our active generative design (AGD) framework can perform an efficient guided search in the design space to identify informative samples for building screening ML models with higher accuracy compared to randomly sampled points. The ML screening model is much more computationally efficient compared to the conventional DFT-based screening model. We have successfully applied our AGD approach to the inverse design of materials with high band gap value and semiconductors and found several potential new materials, including $SrYF_5$, $RbSr_2ClF_4$, and seven other semiconductors, as validated using DFT-based principles. Thus, our approach is expected to be applicable to finding materials with other properties such as thermal conductivity or superionic conductors.

AUTHOR INFORMATION

Corresponding Author

Jianjun Hu − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0002-8725-6660; Email: jianjunh@cse.sc.edu

Authors

Rui Xin − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Edirisuriya M. D. Siriwardane − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0001-8960-5273

Yuqi Song − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Yong Zhao − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0002-6762-266X

Steph-Yves Louis − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Alireza Nasiri − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.1c02438

Author Contributions

Conceptualization: J.H.; methodology: R.X. and J.H.; software: R.X.; validation: R.X., J.H.; investigation: R.X., J.H., E.M.D.S., Y.S., Y.Z., A.N., S.-Y.L.; resources: J.H.; writing—original draft preparation: R.X., J.H.; writing—review and editing: R.X. and J.H.; visualization: R.X. and Y.S.; supervision: J.H.; and funding acquisition: J.H.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research reported in this work was supported in part by NSF under the grant and 1940099 and 1905775 and by NSF SC EPSCoR Program under award number (NSF Award OIA-1655740 and GEAR-CRP 19-GC02). The views, perspectives, and content do not necessarily represent the official views of the National Science Foundation.

REFERENCES

(1) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 3814−3822.

(2) Heo, S. H.; Jo, S.; Kim, H. S.; Choi, G.; Song, J. Y.; Kang, J.-Y.; Park, N.-J.; Ban, H. W.; Kim, F.; Jeong, H.; et al. Composition change-driven texturing and doping in solution-processed SnSe thermoelectric thin films. *Nat. Commun.* **2019**, *10*, No. 864.

(3) Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **2019**, *4*, 331−348.

(4) Noh, J.; Gu, G. H.; Kim, S.; Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **2020**, *11*, 4871−4881.

(5) Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* **2020**, *6*, No. 84.

(6) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, No. 011002.

(7) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1*, No. 15010.

(8) Min, K.; Cho, E. Accelerated discovery of potential ferroelectric perovskite via active learning. *J. Mater. Chem. C* **2020**, 7866−7872.

(9) Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2018**, *2*, No. 0121.

(10) Tagade, P. M.; Adiga, S. P.; Pandian, S.; Park, M. S.; Hariharan, K. S.; Kolake, S. M. Attribute driven inverse materials design using deep learning Bayesian framework. *npj Comput. Mater.* **2019**, *5*, No. 127.
(11) Chen, C.-T.; Gu, G. X. Generative deep neural networks for inverse materials design using backpropagation and active learning. *Adv. Sci.* **2020**, *7*, No. 1902607.
(12) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse design of solid-state materials via a continuous representation. *Matter* **2019**, *1*, 1370−1384.
(13) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **2020**, *6*, No. eaax9324.
(14) Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.* **2018**, *4*, No. 67.
(15) Chen, M.; Jiang, J.; Fan, J. A. Design space reparameterization enforces hard geometric constraints in inverse-designed nanophotonic devices. *ACS Photonics* **2020**, *7*, 3141−3151.
(16) Jiang, J.; Fan, J. A. Multiobjective and categorical global optimization of photonic structures based on ResNet generative neural networks. *Nanophotonics* **2020**, 361−369.
(17) Zhang, Y.; Wang, H.; Wang, Y.; Zhang, L.; Ma, Y. Computer-assisted inverse design of inorganic electrides. *Phys. Rev. X* **2017**, *7*, No. 011017.
(18) Elsawy, M. M.; Lanteri, S.; Duvigneau, R.; Fan, J. A.; Genevet, P. Numerical Optimization Methods for Metasurfaces. *Laser Photonics Rev.* **2020**, *14*, No. 1900445.
(19) Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated discovery of large electrostrains in BaTiO3-based piezoelectrics using active learning. *Adv. Mater.* **2018**, *30*, No. 1702884.
(20) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **2019**, *5*, No. 21.
(21) Settles, B. *Active Learning Literature Survey*; University of Wisconsin-Madison, 2009.
(22) Schröder, C.; Niekler, A. A Survey of Active Learning for Text Classification Using Deep. 2020, arXiv:2008.07267. arXiv.org e-Print archive. https://arxiv.org/abs/2008.07267 (accessed Aug 17, 2020).
(23) Lindenbaum, M.; Markovitch, S.; Rusakov, D. Selective sampling for nearest neighbor classifiers. *Mach. Learn.* **2004**, *54*, 125−152.
(24) Sinha, S.; Ebrahimi, S.; Darrell, T. In *Variational Adversarial Active Learning*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp 5972−5981.
(25) Melville, P.; Mooney, R. J. Diverse ensembles for active learning. *Mach. Learn.* **2004**, 74.
(26) Allahyari, Z.; Oganov, A. R. Coevolutionary search for optimal materials in the space of all possible compounds. *npj Comput. Mater.* **2020**, *6*, No. 55.
(27) Louis, S.-Y.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; Hu, J. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18141−18148.
(28) Goodall, R. E.; Lee, A. A. Predicting Materials Properties Without Crystal Structure: Deep Representation Learning from Stoichiometry. 2019, arXiv:1910.00617. arXiv.org e-Print archive. https://arxiv.org/abs/1910.00617 (accessed Oct 1, 2019).
(29) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95−98.
(30) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, No. 145301.
(31) Snoek, J.; Larochelle, H.; Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2951−2959.
(32) Mockus, J.; Tiesis, V.; Zilinskas, A. The application of Bayesian methods for seeking the extremum. *Towards Global Optim.* **1978**, *2*, 2.
(33) Srinivas, N.; Krause, A.; Kakade, S. M.; Seeger, M. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. 2009, arXiv:0912.3995. arXiv.org e-Print archive. https://arxiv.org/abs/0912.3995 (accessed Dec 21, 2009).
(34) Nogueira, F. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014. https://github.com/fmfn/BayesianOptimization.
(35) Hargreaves, C. J.; Dyer, M. S.; Gaultois, M. W.; Kurlin, V. A.; Rosseinsky, M. J. The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions. *Chem. Mater.* **2020**, 10610−10620.
(36) Kresse, G.; Hafner, J. Ab initiomolecular dynamics for liquid metals. *Phys. Rev. B* **1993**, *47*, 558−561.
(37) Kresse, G.; Hafner, J. Ab initiomolecular-dynamics simulation of the liquid-metal−amorphous-semiconductor transition in germanium. *Phys. Rev. B* **1994**, *49*, 14251−14269.
(38) Kresse, G.; Furthmüller, J. Efficiency of ab initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(39) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for ab initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169−11186.
(40) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953−17979.
(41) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758−1775.
(42) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(43) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396.
(44) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(45) Hill, R. The Elastic Behaviour of a Crystalline Aggregate. *Proc. Phys. Soc., Sect. A* **1952**, *65*, 349−354.
(46) Mouhat, F.; Coudert, F.-X. Necessary and Sufficient Elastic Stability Conditions in Various Crystal Systems. *Phys. Rev. B* **2014**, *90*, No. 224104.
(47) Wang, V.; Xu, N.; Liu, J.; Tang, G.; Geng, W. VASPKIT: A User-friendly Interface Facilitating High-throughput Computing and Analysis Using VASP Code. *Comput. Phys. Commun.* **2021**, No. 108033.
(48) Xiong, Z.; Cui, Y.; Liu, Z.; Zhao, Y.; Hu, M.; Hu, J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput. Mater. Sci.* **2020**, *171*, No. 109203.
(49) Loftis, C.; Yuan, K.; Zhao, Y.; Hu, M.; Hu, J. Lattice Thermal Conductivity Prediction Using Symbolic Regression and Machine Learning. *J. Phys. Chem. A* **2020**, 435−450.