THE JOURNAL OF
PHYSICAL CHEMISTRY C

Accelerated Discovery of Novel Inorganic Materials with Desired Properties Using Active Learning

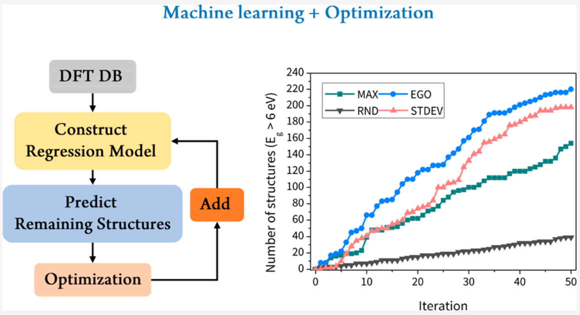
Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Kyoungmin Min* and Eunseog Cho*

Cite This: *J. Phys. Chem. C* 2020, 124, 14759−14767 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Construction of prediction models using machine learning algorithms on existing databases expands the search limit of undiscovered structures, in principle, to the entire materials space. However, because of uncertainties in machine learning prediction, the suggested properties are not always promising; thus, improving the database quality is mandatory for validation as well as improvement in prediction accuracy. To achieve this, we herein implement an active learning process, beginning with a limited number of databases, to find materials satisfying target properties (band gap and refractive index) with minimized trials and errors. The regression model is initially trained with only around 2% of the entire search space, and 20 new databases, suggested from the optimization schemes, are added at each optimization process. Between exploration, exploitation, random selection, and the Bayesian optimization method, the Bayesian method exhibits the best performance in finding the number of materials that satisfies the criteria within limited trials In addition, the structure with the maximum target property values is found after searching only around 7.0% and 7.7% of the entire database for band gap and refractive index, respectively. Current results clearly confirm that the active learning process can be accelerated to find ideal materials satisfying target properties with minimized resources.



■ INTRODUCTION

In many materials science fields, a high-throughput computational screening method has been demonstrated to design novel materials satisfying target properties by exploring all possible combinations in a material composition.[1−3] This is possible due to rapid advances in computing power and resources as well as the development of accurate calculation methods based on *ab initio* approaches. However, applying a materials screening process to a vast amount of chemical space still requires a lot of resources, limiting their extensive applications. The advent of machine learning (ML) algorithms has been anticipated as a way to resolve this issue because based on previously explored chemical space the machine can mimic the function to characterize the structure (or material) to properties relations. Then, the trained surrogate model can be used to predict the properties of unexplored materials with the speed of 100 orders of magnitude faster than conventional but novel calculation methods: density functional theories (DFTs).

Many successful implementations of ML approaches to materials science have been demonstrated, such as predicting the band gap, formation energy, elastic constant, and other fundamental materials properties.[4−8] However, extrapolation from the prediction model remains an unresolved problem in the ML field. Usually, training set prediction accuracy in ML-based models is reasonable as long as there is a practically adequate number of training sets. However, this does not mean that all possible chemical spaces from the pre-existing database are covered. Hence, it is unfortunate that when predicting properties of newly explored structures it is highly probable that prediction reliability is beyond the previously obtained range of error. This problem could be partially resolved by simply including more databases, but this is not practical and largely inefficient because it is unclear which information should be added from all possible chemical spaces in materials.

In this regard, it is important to suggest the direction of the materials search by including an optimization process to decide which database should be added first to find target materials efficiently. The conventional optimization process is performed to satisfy the following two purposes: (1) to decrease

Received: January 20, 2020
Revised: June 5, 2020
Published: June 28, 2020

https://dx.doi.org/10.1021/acs.jpcc.0c00545

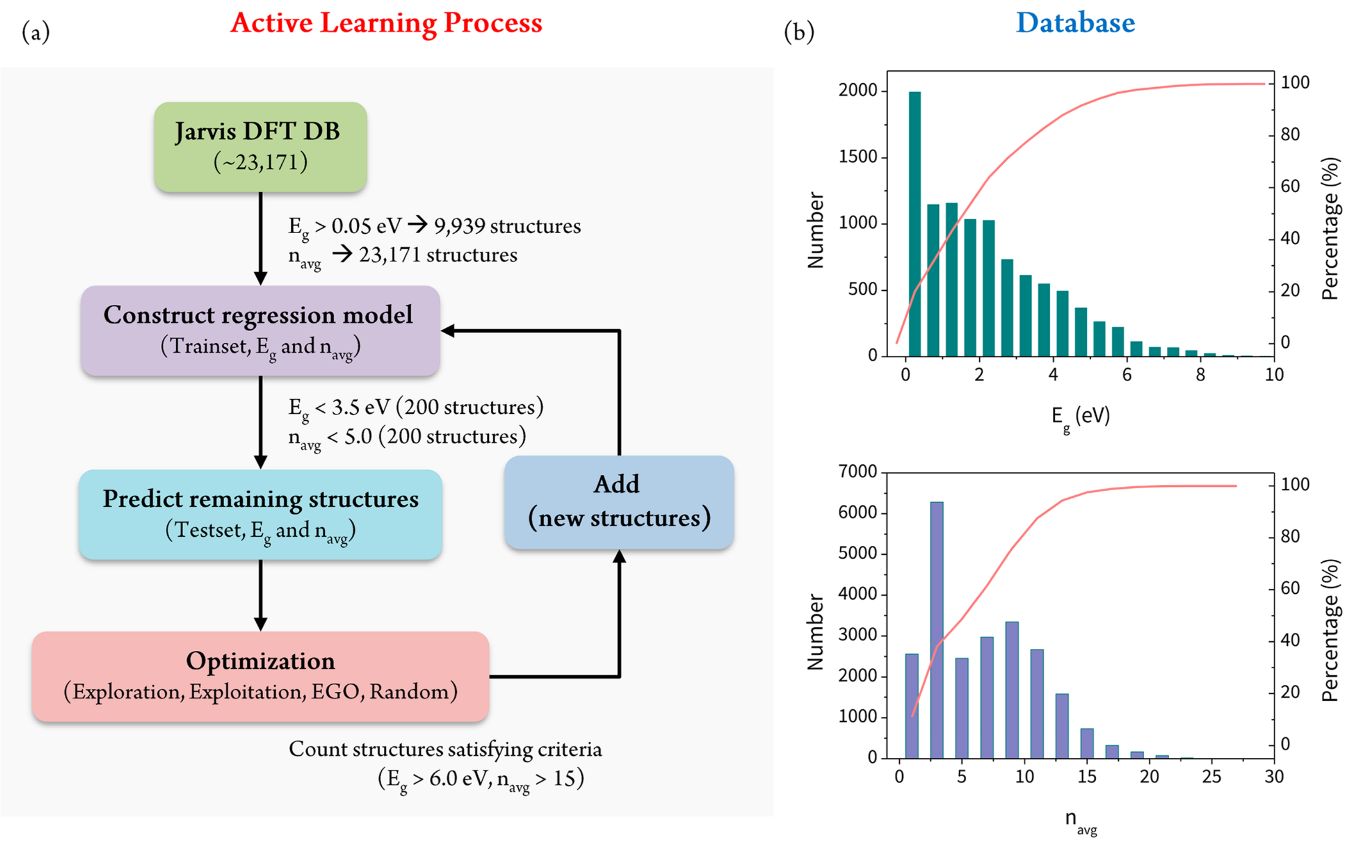


Figure 1. (a) Schematic for the active learning process. (b) Entire database for (top) band gap (E_g) and (bottom) refractive index (n_{avg}). Red line in (b) indicates the cumulative percentage of each property.

uncertainty in the prediction accuracy (exploration) and (2) to propose the best materials based on the current prediction model (exploitation). By the exploration method, prediction accuracy can be increased by adding a database whose predicted values exhibit low confidence. On the other hand, the exploitation method is particularly useful when the prediction model already has reasonable accuracy and the error range of the predicted values is expected to be comparable to that of the training values. However, it is critical to satisfy both conditions, i.e., to decrease uncertainty in the prediction model while simultaneously suggesting an ideal structure with the target properties. To meet these criteria, efficient global optimization (EGO) has been suggested as a compromise between exploitation and exploration to avoid allowing the prediction model to fall into the local minima due to adding a single-purpose database.[9] This method calculates the expected improvement (EI) value for the list of candidate materials; by using it, one can choose which of them should be considered first in order to find materials with a larger or a smaller target property value than the current optimal value. The EGO algorithm has outperformed other optimization methods in several cases such as finding high-temperature ferroelectric perovskites,[10] optimal layered materials,[11] and new piezoelectrics with large electrostrains.[12]

In this study, we demonstrate the performance of the active learning process for practical application in inorganic solid materials to find structures with larger band gaps or refractive indices. Knowing the band gap value is a fundamental step in determining whether the candidate material is metal, semiconductor, or insulator; additionally, finding materials with larger band gaps is often required to design novel electronic devices that prevent current leakage or optoelectronic devices where transparency is necessary and switching at larger voltage ranges is crucial.[13,14] Materials with a high refractive index are also widely used in optical applications such as optical fibers, antireflective coatings, and optical sensors.[15,16] We begin with a small database, which is randomly chosen from existing inorganic databases, and train it with an ML algorithm to construct a prediction model for each property. Then, three different optimization algorithms (exploration, exploitation, and EGO) are applied, and their performance, including random selection, is compared to validate which of them can find more structures satisfying the criteria with limited addition of new structures in the training set. Finally, the underlying mechanism and the importance of using the active learning process will be discussed.

■ METHODS

Active Learning Process. The schematic view on how the active learning process is implemented is shown in Figure 1(a). This describes the closed-loop process, which gives iterative feedback to the prediction model by adding an informative database, recommended from the optimization method. We first obtain structures from the Jarvis-NIST database,[17] which implements density functional calculations for obtaining the band gap (E_g) and refractive index (n_{avg}) for the majority of inorganic materials (a total of 23 171). Among them, the structures whose E_g is larger than 0.05 eV are extracted, leading to another set of databases with a total of 9939 structures. These database sets are assumed to be whole inorganic chemical spaces to be explored for validation of the active learning process in this

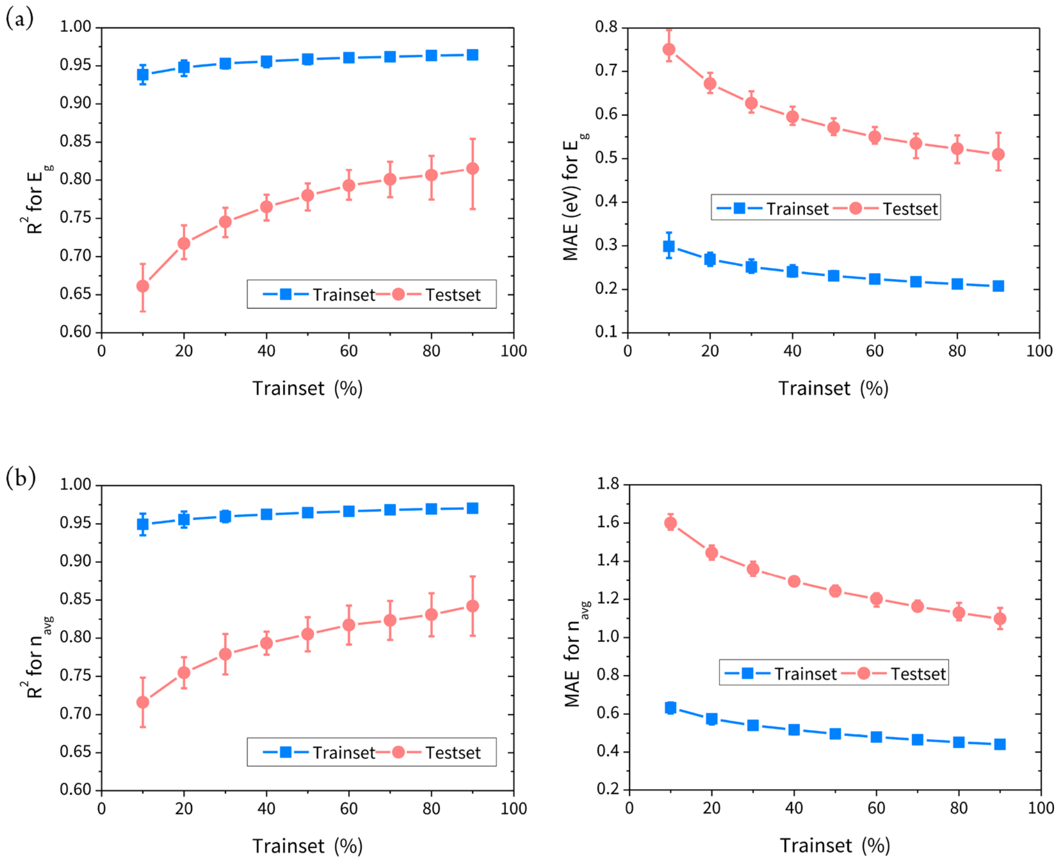


Figure 2. (Left) Coefficient of determinant (R^2) and (right) mean absolute error (MAE) for the test set from the prediction model for (a) band gap (E_g) and (b) refractive index (n_{avg}) depending on the size of the training set. The error bar indicates the standard deviation value.

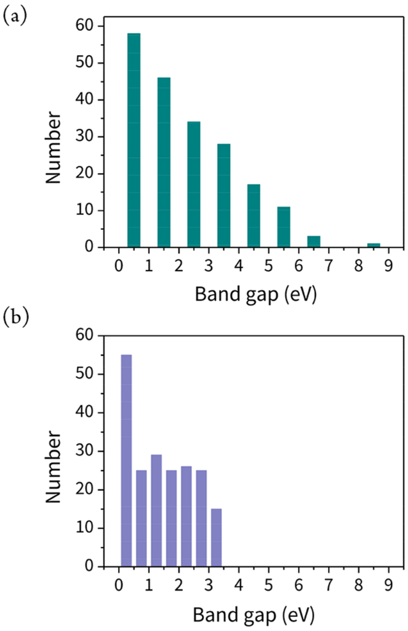
study. An overview of the entire database is shown in Figure 1(b).

To validate the practical functionality of active learning, the initial regression model for E_g and n_{avg} is constructed based on only 200 databases whose values are E_g < 3.5 eV or n_{avg} < 5.0, which are randomly selected from the entire database. We note that more than 50% of the total database matches these criteria. With this small number of databases, one can validate the practical capability of the active learning process because scarce and imbalanced databases are a common problem which materials scientists need to overcome when constructing ML-trained models. Based on the surrogate model in the training set, E_g and n_{avg} are predicted for the rest of the test set. Among the predicted samples, 20 new structures, suggested by each optimization algorithm, are added to the training set, and candidate structures whose E_g > 6.0 eV (3.4% in total) or n_{avg} > 15.0 (5.6% in total) are counted cumulatively during iteration. This loop is first iterated up to 50 times (with the addition of a total of 1000 new structures) and continued unless the structure with the maximum value is found.

Machine Learning and Optimization Methods. In order to construct the ML surrogate model, a random forest (RF) regressor is chosen due to its versatility, and Scikit-learn, a Python ML package, is used for its implementation.[18] For hyperparameter tuning, a randomized search algorithm is performed to find their optimal configuration. The regression model is constructed based on a training set of 90% of the initial database (initially, 200 structures are used, and then 20 recommended structures from the optimization process are added to the training set during the active learning loop), and the remaining 10% is used as a validation set. For cross-validation, the data set is split in a random fashion. Fifty different prediction models are constructed at each iteration of the active learning process based on 90% of the randomly selected training set, in order to obtain statistical outputs (average and standard deviation) in the models. For material descriptors to represent the structures, we implement 145 common chemical features developed by Ward et al.[4] In addition to those, the space number for each structure is added to distinguish the polymorph. It is important to mention that adding more descriptors such as structural features in addition to chemical descriptors and the space group could improve the prediction accuracy. A previous reference[19] shows that using all attributes (chemical and structural descriptors) decreases the error in prediction of the formation energy by 7% for the OQMD database. However, they also mention that for the ICSD training set the prediction accuracy is approximately the same regardless including structural descriptors. In this regard, using chemical descriptors and the space group as input features should be accurate enough, and this implementation can be justified by thr following reasons.

(1) The prediction accuracy for E_g with chemical descriptors is around 0.80 (R^2) with 0.49 eV (MAE), which is reasonably accurate. This error comes mostly from structures with a smaller value of E_g (<2 eV), as shown in Figure S1, Supporting Information (SI). Hence, it will not significantly affect finding the structure with a larger E_g value (>6.0 eV).

(2) As mentioned previously,[19] the prediction accuracy for each model depends more on the database configuration.



	R^2	MAE (eV)
Average	0.504	0.951
Standard dev.	0.032	0.033
Max	0.568	1.040
Min	0.406	0.881

	R^2	MAE (eV)
Average	0.064	1.249
Standard dev.	0.036	0.022
Max	0.160	1.287
Min	0.002	1.183

Figure 3. (Left) Band gap values used to construct the prediction model and (right) corresponding R^2 and MAE values and their statistical output (from 20 different prediction models from a randomly chosen training set) for the test set when (a) the database is chosen randomly from the entire band gap database and (b) the band gap is less than 3.5 eV.

Including more descriptors does not necessarily mean that they improve the prediction accuracy significantly.

(3) The purpose of the current study is to validate the practical applicability of the active learning platform. Since the prediction model works reasonably well, we focus more on comparing the performance of optimization schemes and their potential problems. In other words, "if the prediction accuracy of the current method was not enough, the proposed active learning platform would have not worked efficiently".

(4) Although the space number is not directly related to the electronic properties, including such a descriptor slightly improves the prediction accuracy of E_g (R^2: 0.806 vs 0.800 and MAE: 0.523 eV vs 0.535 eV with and without space number, respectively, when 80% of the train set is used). The space number would not play an important role when the number of training databases is very small during several steps of the initial active learning process because the number of polymorph structures is small. However, its contribution will be gradually increased as more data sets are added (more polymorphs) for training.

For the optimization process, we compare the performance of four different methods: exploration using standard deviation (STDEV), exploitation using maximum (MAX), efficient global optimization (EGO), and random (RND) selection (which is performed for reference). First, exploration is conducted by STDEV, which involves adding structures to the training set during the active learning loop whose predicted values in the test set exhibit the largest deviation depending on 50 different prediction models. This method is particularly useful when the capability of the initial prediction model is poor because predicting uncertainty can be largely reduced using exploration. Meanwhile, the exploitation method MAX is used to recommend structures with the largest average predicted value out of 50 prediction models. Implementing this method is recommended when the accuracy of the prediction model is already good enough to propose the structures with ideal properties. Lastly, the Bayesian optimization method, EGO, is performed with the acquisition function of the expected improvement (EI). This method works efficiently by preventing bias in the direction of data addition by a compromise between exploitation and exploration. EGO works by recommending structures whose EI values are the maximum. EI is defined as[9]

$$\mathrm{EI}(x) = (\mu(x) - \mu_{\max})\Phi(z) + \sigma(x)\phi(z)$$

where $z = (\mu(x) - \mu_{\max} - \varepsilon)/\sigma(x)$; $\mu(x)$ and $\sigma(x)$ are the predicted mean and standard deviation, respectively, for a given material x from the RF regressor; $\mu_{\max}$ is the largest value observed thus far in the training database; $\Phi(z)$ is the standard cumulative distribution function; $\phi(z)$ is the standard normal density function; and ε is the trade-off parameter. 0.01 is used for ε.

■ RESULTS AND DISCUSSION

Prediction Model Capability. Based on the entire Jarvis-NIST database, we first constructed the prediction model for the individual properties of E_g and n_{avg} to investigate the sensitivity of prediction uncertainty depending on various test-to-training set ratios. As shown in Figure 2, in general, it shows that a smaller training set results in poorer performance (smaller coefficient of determinant (R^2) and larger mean absolute error (MAE)) on the test set prediction, as expected. As an example, detailed results on how the test set vs predicted values are distributed, and their corresponding error values are shown in Figure S1 (Supporting Information) when the data are split to 8:2 for the training and test sets, respectively. It clearly validates that the ML prediction model is well-constructed (R^2 > 0.84 for both cases), and more than 75% of the error lies between −0.5 to 0.5 eV and −1.5 to 1.5 for E_g and n_{avg}, respectively.

In Figure 2(a), it is seen that from a 10% to 90% increase in the training set range the obtained R^2 value starts with an unexpectedly large value of around 0.65∼0.72 and increases to over 0.8 for both quantities. In other words, this means that when trained with only 10% of the total database the constructed prediction model can afford to be used as a coarse screening tool since the trained model can still predict the remaining 90% with small MAE values of 0.75 eV and 1.6 for E_g and n_{avg}, respectively. This is possible because when randomly selecting the training set from the whole database structures with larger target properties are also likely to be included, and so the training region is not limited to a certain range of values. Then, the possible risk of extrapolation, to which the ML model is the most vulnerable, could be reduced, leading to better prediction accuracy.

Database Configuration for Active Learning. When practically dealing with databases in scientific fields, the values in a constructed database are not always evenly distributed, and the amount of data may not be large enough to represent the chemical space of interest. Rather, the number of databases is often limited, and data distribution is biased and imbalanced, which makes it difficult to construct a reliable prediction model. In this regard, to confirm the practical functionality of implementing active learning, we hypothesize that the initial number of databases is only 200 (2% and 0.9% out of the total database for E_g and n_{avg}, respectively), and all of them have a value of less than 3.5 eV for E_g (or 5 for n_{avg}). In this case, the initial prediction model could work reasonably well within the trained region but not out of it (extrapolation).

To confirm this hypothesis, the performance of two different models predicting E_g is compared when the training database is configured from the whole database (Figure 3(a), Case A) or from the database whose E_g is less than 3.5 eV (Figure 3(b), Case B). Structures are randomly selected, and the distribution of the database is shown on the left in Figure 3. Among the databases, 90% and 10% are chosen for the training and validation set, respectively, and the remaining (9939−200 = 9739) structures are predicted (test set). 90% of the training set is randomly selected 20 times, and the prediction model is constructed to obtain the statistical output for each case. On the right in Figure 3, the prediction accuracy for each model shows that Case A exhibits moderate performance (R^2 = 0.5 and MAE = 0.95 eV for the test set). However, performance for Case B indicates that now its R^2 (0.064) and MAE (1.249 eV) become much poorer. To provide a more detailed analysis, the predicted vs test set and its corresponding error are shown in Figure S2, Supporting Information. First, it shows that the maximum predicted value from Case A is around 6 eV, although the covered range of the test set is up to around 10 eV (Figure S2(a), Supporting Information). This result becomes even poorer for Case B (Figure S2(b), Supporting Information), where none of the predicted E_g values exceed 3 eV; thus, the maximum error is 8.06 eV. This clearly confirms that when the surrogate model predicts properties of structures outside of the training range the predicted values hardly provide reliable accuracy, so the importance of implementing active learning becomes critical.

Considering the results for Case B, we demonstrate how the active learning process can discover the structure with the largest E_g or n_{avg} with a limited number of iterations. We also establish how many structures satisfying the given criteria (larger than 6 eV or 15 for E_g or n_{avg}, respectively) are found. For reference, the material whose value is the largest in the initial database configuration and the list of the top five target materials are shown in Table S1, Supporting Information.

Performance of Active Learning. The performance of the active learning process implementation is demonstrated first for the E_g case. During iteration, the cumulative number of structures satisfying the criterion E_g > 6 eV (a total of 339 structures, which is only 3.4% of the entire database) is counted for each optimization method as shown in Figure 4(a). First, it clearly shows that all the optimization schemes perform much better than the random search, and among them, EGO works the most efficiently. After 50 iterations, the total number of satisfying structures from EGO reaches 220, which is 64% of the 339 total existing structures whose E_g is larger than 6 eV (154, 198, and 39 structures found using MAX, STDEV, and RND, respectively). On average, 12 out of 20 (the number of databases added at each iteration) satisfy the criterion when using EGO. This result clearly confirms that EGO can be greatly useful in using minimum trials to find the materials which possess the target properties. In addition, it is interesting to note that the MAX model suggests the materials whose E_g is larger than 3.5 eV, although those materials are not included in the training set. This is because the MAX model recommends the materials whose predicted value is the largest, and the predicted E_g for materials with actually a larger value is not always smaller than the others. During the initial few steps of active learning, 0∼2 number of materials are recommended from MAX, but they expand the prediction range of the training set gradually; hence, the efficiency from the MAX model will be enhanced eventually.

Furthermore, it is worth noting that the STDEV exploration method still performs reasonably well. This could be attributed

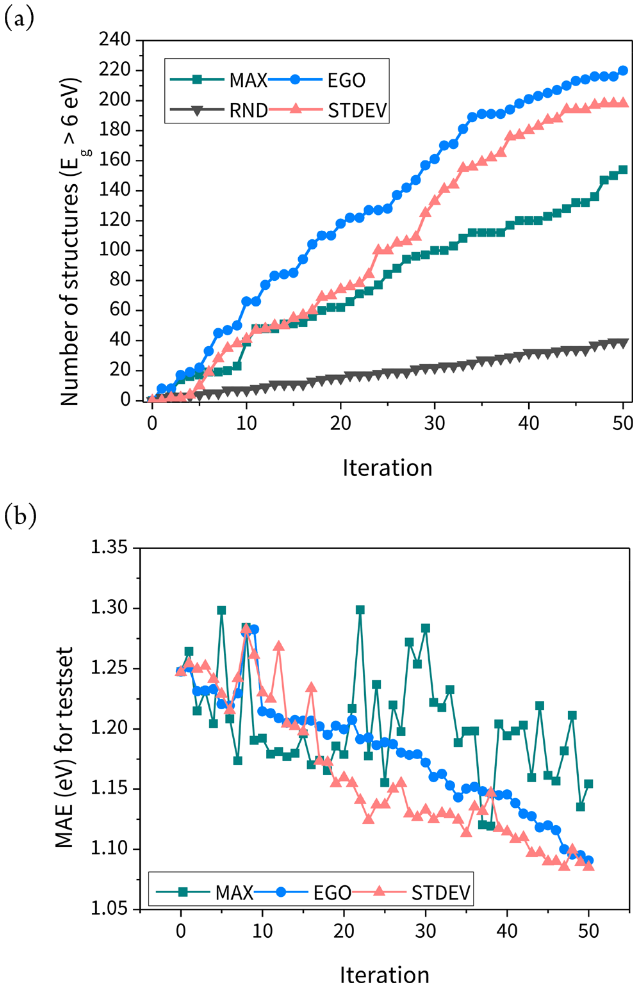


Figure 4. (a) Cumulative number of structures satisfying the criterion E_g > 6 eV and (b) MAE for the test set during active learning. Twenty new data sets are added at each iteration.

to an increase in prediction accuracy because this method keeps suggesting structures with the largest prediction uncertainty. As shown in Figure 4(b), because the training database increasingly includes those structures, its prediction accuracy continues to grow faster here than in the other schemes. It confirms that the MAE value from STDEV noticeably decreases more than others so that the trained ML model can capture the structures with larger E_g values. MAE from the other two schemes (EGO and MAX) also decreases because the training database is gradually expanding during iterations. We also note that sudden peaks observed in MAE, especially from the MAX results, are due to training overfitting because we predict values in the test set which are much larger than those in the training set. This overfitting is scarcely observed from the EGO and STDEV results because, unlike MAX, they consider the prediction uncertainty of the database.

The outperformance from EGO is expected to some extent because, as mentioned in the Introduction, its efficiency has already been demonstrated in previous cases,[10−12] and it compromises between exploration and exploitation by considering both accuracy and uncertainty.[9] However, the remarkable functionality of the others exceeds our expectations, especially the MAX exploitation method, which shows unexpectedly good performance; the cumulative number of satisfying structures reaches 154 (45% of the possible structures) in Figure 4(a). In this regard, we obtained the average standard deviation as well as the average values of the predicted E_g for two groups: Group A, structures with $E_g > 3.5$ eV, and Group B, $E_g < 3.5$ eV, as shown in Figure S3, Supporting Information. It is important that predicted values in Group A always exhibit a larger standard deviation than those in Group B (Figure S3(a), Supporting Information). In addition, its predicted values are always larger than those in Group B, although there is a clear difference between predicted and actual values, as shown in Figure S3(b), Supporting Information. This means that during optimization structures in Group A have a much greater possibility of being recommended by the STDEV and MAX scheme, leading to suggest more structures satisfying targeted properties than random sampling.

For further investigation of initial database dependency, the performance of MAX, STDEV, and EGO, when the number of the training sets is reduced to 100 and 50, is compared. Basically, in Figure S4 (Supporting Information), it shows that the initial performance for smaller training sets is poorer. However, when the number of training sets reaches 200 for both cases (after 5 and 8 iterations for the number of training sets with 100 and 50, respectively), the performance of all optimization methods becomes comparable to the case when the number of training sets is initially 200 (the slope in the curve is similar). This is because with the same number of training sets the overall composition in E_g value is similar to all cases.

In addition to the demonstration for E_g, we further examine the performance of the active learning process for finding structures with a larger n_{avg} value. Since a broader chemical space needs to be explored for n_{avg} than for E_g (23 171 vs 9 939 structures, respectively), increased difficulty in searching for structures satisfying the criterion is expected. The criterion is set as $n_{avg} > 15$, comprised of 1309 structures, which is only 5.65% of the entire database.

First, as discussed in the case of E_g, the cumulative number of structures satisfying the target property is obtained during iteration for all optimization schemes shown in Figure 5(a). Again, all the three methods outperform the random search, as expected, and among them, the MAX method works most efficiently by finding 236 structures, which is 18% of the total number of structures satisfying the criterion. The EGO method follows in terms of efficiency, finding a total of 14% of the structures that satisfy the criterion; STDEV also shows acceptable performance, finding 11% in total. It is also worth noting that among the suggested structures (1000 in total during 50 loops) the number of structures recommended by MAX, EGO, STDEV, and RND whose n_{avg} value is larger than 4.70 (the largest initial n_{avg} value) is 947, 883, 871, and 232, respectively, confirming the efficiency of the optimization methods. In terms of the MAE change during iteration, as similarly observed in the E_g case, the MAE value in the test set keeps decreasing due to an increase in the number of databases in the training set (Figure 5(b)), and STDEV exhibits the best accuracy in general.

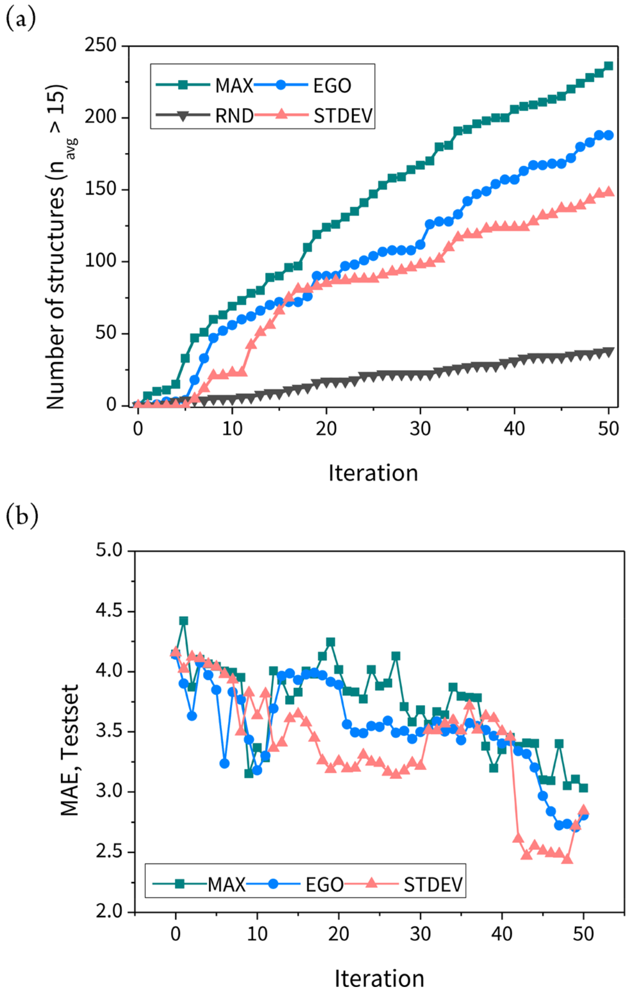


Figure 5. (a) Cumulative number of structures found satisfying the criterion $n_{avg} > 15$ and (b) MAE for the test set during active learning. Twenty new data sets are added at each iteration.

The overall performance and comparison of optimization methods and target properties are summarized in Table 1. The average number (selection rate) of structures found at each iteration out of the 20 suggested structures clearly demonstrates the superior performance (more than three structures are found on average) of the optimization process compared to that of the random selection (less than one structure is found). It is also important that the optimization process can help to find the structure with the maximum value in the database. For example,

Table 1. Overall Performance of Each Active Learning Process for the Band Gap (E_g) and Refractive Index (n_{avg})[a]

	E_g				n_{avg}			
category	MAX	EGO	STDEV	RND	MAX	EGO	STDEV	RND
average number	3.0	4.4	3.9	0.78 ± 0.08	4.7	3.7	2.9	0.76 ± 0.10
selection rate	15.4%	22%	19.8%	3.9% ± 0.4%	23.6%	18.8%	14.8%	3.8% ± 0.5%
cumulative	154 (45%)	220 (64%)	198 (58%)	39 ± 4 (12% ± 1)	236 (18%)	188 (14%)	148 (11%)	38 ± 5 (3% ± 0.3)
max. value	9.75 eV (1st)	9.75 eV (1st)	9.63 eV (3rd)	n/a	24.15 (3rd)	24.33 (2nd)	20.51 (13th)	n/a
max. value found at	31th	25th	68th	n/a	85th	80th	98th	n/a

[a] The number in the parentheses at cumulative indicates the percentage of found materials among the satisfying materials.

both EGO and MAX discover the structure with the maximum E_g value (9.75 eV) at the 25th and 31st iteration, respectively. Meanwhile, STDEV finds the third largest structure within 50 iterations and finally suggests the best material at the 68th iteration.

In the case of n_{avg}, EGO and MAX find the second and third largest structures in the entire database within 50 iterations, while STDEV exhibits slightly poorer performance with finding the 13th largest structure. EGO, MAX, and STDEV discover the structure with the maximum value at the 80th, 85th, and 98th iteration, respectively. Based on these results, we conclude that, in general, EGO is the most effective method based on the fact that (1) the structures with the largest values of both E_g and n_{avg} are suggested at the earliest, after searching only ∼8% of the entire database and (2) the cumulative number of structures satisfying the criterion is the largest for E_g and the second largest for n_{avg}.

Furthermore, to validate the current trade-off value (0.01), ε is increased to 0.05 and 0.1 (more exploration), and then the performance of EGO is compared. As shown in Figure S5 (Supporting Information), the maximum number of structures satisfying the target conditions becomes less as ε is increased for both E_g and n_{avg}. This could be attributed to the fact that STDEV works more poorly than the current setup ($\varepsilon = 0.01$) of EGO, and thus more exploration is not effective for improving the performance of EGO. It is anticipated that for the case when the entire chemical space is much larger than the current problem ($>10^5$) more exploration would help increase the prediction accuracy.

It would be arguable that the data set used in this study (Jarvis-NIST) is not accurate enough to represent E_g and n_{avg} of inorganic materials. This is partially true because that database did not consider the intraband transition, and it leads to omit the Drude-like transition when calculating metallic systems.[17,20] Hence, their dielectric data are more reliable for materials with high E_g value,[21] and the additional step to compute the plasma frequency needs to be conducted for improved accuracy.[22] In this regard, the extra concern should be made to finalize the materials search with current active learning platform since the final target materials with a large value of n_{avg} in Table S1 (Supporting Information) are metallic. Hence, further calculation should be conducted for final validation of the suggested materials when implementing this platform for finding materials with high refractive index in practice.

Nonetheless, this critical issue does not devaluate the core findings of current work for the following reasons. First, we confirm that the current machine learning model shows great performance for predicting DFT results; thus, predicting the contribution from Drude-like transition is expected to be also accurate when those data are available. In addition, when the purpose of implementing the active learning is to find the materials with large n_{avg} as well as E_g, one could ignore calculating the n_{avg} for structures with small E_g value. Second, the purpose of the current work is mainly to validate the practical functionality of the active learning platform for a materials screening process. In this regard, the current study focuses not only on the performance for finding the best materials but also validating if the materials, whose properties are at least better than the currently available material, could be suggested. Then, this proves that the active learning platform exhibits superior performance than the random search method. In addition, it is also beneficial to validate that the current active learning platform works on the well-established calculation platform (DFT) because researchers will use this methodology anyway when they look for the best materials. In this regard, instead of considering optimization of the calculation method itself, we focus more on validating the performance of the active learning platform. It is also important to note that in Figure 5 the cumulative number of satisfying structures ($n_{avg} > 15$) is predominantly more from the active learning than the random search, clearly indicating that the current method is quite practical and functional for finding the target structures.

Reasoning for Implementing Active Learning. By using descriptors to group the structures with respect to their properties, a deeper understanding of why the active learning process is a crucial step in searching for target structures could be achieved. If structures with similar properties are accumulated in the same region, then the material scientist could search for structures specific to the region where the target property is located. To validate this scenario, the dimensions of ML features must be reduced because the current number of descriptors (146) cannot be represented in a two-dimensional space. We therefore implement the t-distributed stochastic neighbor embedding (t-SNE) algorithm to reduce the high-dimensional data set to a two-dimensional space for clustering.[23] For the parameters setup, the Barnes−Hut algorithm is used for faster performance, and Euclidean distance is calculated for distance metrics.

After clustering with t-SNE as shown in Figure 6, it shows that structures satisfying the criterion (blue circles) are unexpectedly not accumulated in the space for both E_g and n_{avg}. Rather, the distribution of structures seems to be random with respect to their properties, indicating that it is not feasible to group the structures, at least with the current descriptors. If structures could be practically clustered, it might have been possible to search the whole chemical space to find the region where the structures with the target values are grouped. Once the target region is found, it would be much easier to suggest candidate structures. This scenario does not apply to the current case, making it even more important to optimize, because it is much more difficult to direct the search for the locations of the target materials. Poor performance from t-SNE implementation would be attributed to weak direct relation between machine learning descriptors and target property. For example, the Pearson

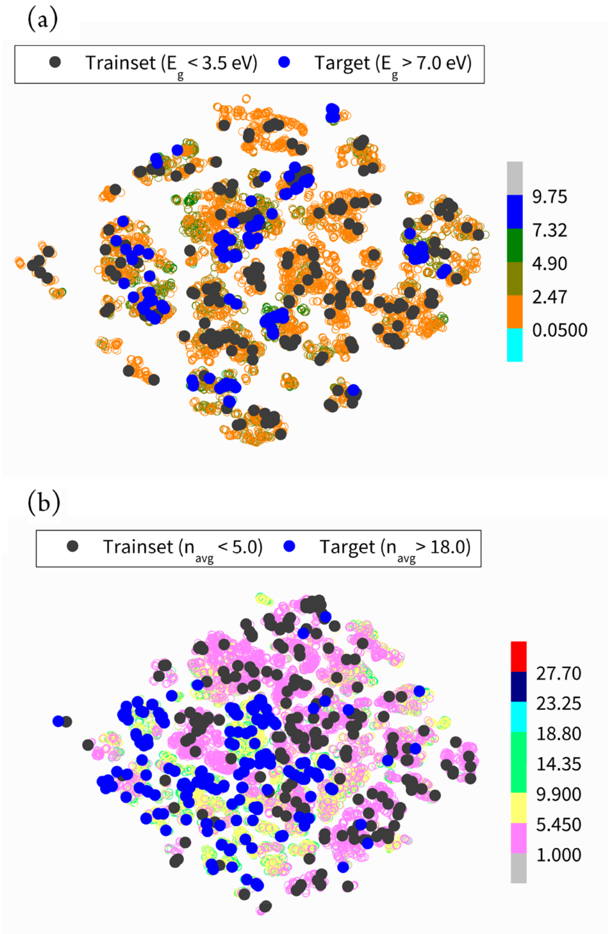


Figure 6. Dimensional reduction and clustering plot using t-SNE for (a) band gap (E_g) and (b) refractive index (n_{avg}).

correlation value shown in Tables S2 and S3 (Supporting Information) indicates that the maximum R is obtained to be −0.411 (frac_dValence) and −0.478 (mean_NpValence) for E_g and n_{avg} prediction. (Details for the description of each feature can be found in ref 4.) In addition, only 5 and 20 descriptors (for E_g and n_{avg}, respectively) exhibit that their absolute R value is larger than 0.35. In this respect, the loss of information and correlation could be accelerated while reducing the dimension from 146 to 2, leading to poor performance in t-SNE.

CONCLUSIONS

In this study, we implement the active learning process to validate its functionality to accelerate materials discovery with target band gap and refractive index properties. Although the ML-based prediction model is initially trained with only 1∼2% of the database of the entire search space, it is surprising that after suggesting 50 loops with a total of 1000 new structures through the Bayesian-based optimization method it finds 64% (220) of the total number of structures whose E_g is larger than 6 eV (339 structures, which is only 3.4% of the entire database) and 14% (188) of the total number of structures whose n_{avg} is larger than 15 (1309 structures, which is only 5.6% of the entire database). The other conventional optimization methods, exploration (STDEV) and exploitation (MAX), also show great performance, but EGO works the most efficiently in terms of suggesting a total number of structures satisfying the target criterion and finding the structure with the largest value. The structure with the maximum E_g or n_{avg} value is found after searching only 7.0% and 7.7% of the total database, respectively. Furthermore, implementing dimensional reduction via unsupervised learning in the clustering method exhibits that the structures with larger values of the given properties are not grouped in the reduced dimension, indicating that the role of the optimization process is much more significant in providing the direction of the materials search. Current results clearly confirm that the active learning process works efficiently to design materials that satisfy targeted properties, which can greatly reduce the need for computational (and potentially experimental) resources.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.0c00545.

Test set prediction error, average standard deviation in the prediction value, properties in the initial database, comparison of performance under various conditions, and Pearson correlation value (PDF)

AUTHOR INFORMATION

Corresponding Authors

Kyoungmin Min − *Autonomous Material Development Lab, Samsung Advanced Institute of Technology, Suwon, Gyeonggi-do 16678, Republic of Korea;* orcid.org/0000-0002-1041-6005; Email: kmin.min@ssu.ac.kr

Eunseog Cho − *Autonomous Material Development Lab, Samsung Advanced Institute of Technology, Suwon, Gyeonggi-do 16678, Republic of Korea;* orcid.org/0000-0001-5308-8278; Email: eunseog.cho@samsung.com

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpcc.0c00545

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Mounet, N.; Gibertini, M.; Schwaller, P.; Campi, D.; Merkys, A.; Marrazzo, A.; Sohier, T.; Castelli, I. E.; Cepellotti, A.; Pizzi, G.; et al. Two-Dimensional Materials from High-Throughput Computational Exfoliation of Experimentally Known Compounds. *Nat. Nanotechnol.* **2018**, *13* (3), 246−252.

(2) Min, K.; Seo, S.-W.; Choi, B.; Park, K.; Cho, E. Computational Screening for Design of Optimal Coating Materials to Suppress Gas Evolution in Li-Ion Battery Cathodes. *ACS Appl. Mater. Interfaces* **2017**, *9* (21), 17822−17834.

(3) Körbel, S.; Marques, M. A. L.; Botti, S. Stability and Electronic Properties of New Inorganic Perovskites from High-Throughput Ab Initio Calculations. *J. Mater. Chem. C* **2016**, *4* (15), 3157−3167.

(4) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *Npj Comput. Mater.* **2016**, *2*, 16028.

(5) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

(6) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. L. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chem. Mater.* **2017**, *29* (12), 5090−5103.

(7) Takahashi, K.; Takahashi, L.; Miyazato, I.; Tanaka, Y. Searching for Hidden Perovskite Materials for Photovoltaic Systems by

Combining Data Science and First Principle Calculations. *ACS Photonics* **2018**, *5* (3), 771−775.

(8) Min, K.; Choi, B.; Park, K.; Cho, E. Machine Learning Assisted Optimization of Electrochemical Properties for Ni-Rich Cathode Materials. *Sci. Rep.* **2018**, *8* (1), 15778.

(9) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **1998**, *13* (4), 455−492.

(10) Balachandran, P. V.; Kowalski, B.; Sehirlioglu, A.; Lookman, T. Experimental Search for High-Temperature Ferroelectric Perovskites Guided by Two-Step Machine Learning. *Nat. Commun.* **2018**, *9* (1), 1668.

(11) Bassman, L.; Rajak, P.; Kalia, R. K.; Nakano, A.; Sha, F.; Sun, J.; Singh, D. J.; Aykol, M.; Huck, P.; Persson, K.; et al. Active Learning for Accelerated Design of Layered Materials. *npj Comput. Mater.* **2018**, *4* (1), 74.

(12) 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* (7), 1702884.

(13) Wilk, G. D.; Wallace, R. M.; Anthony, J. M. High-κ Gate Dielectrics: Current Status and Materials Properties Considerations. *J. Appl. Phys.* **2001**, *89* (10), 5243−5275.

(14) Kirschman, R. K. *High Temperature Electronics*; IEEE press, 1999.

(15) Biron, M. 7 - Plastics Solutions for Practical Problems. In *Plastics Design Library*, 2nd ed.; Biron, M. B. T.-T., Ed.; William Andrew Publishing, 2013; pp 831−984.

(16) Liu, J.; Ueda, M. High Refractive Index Polymers: Fundamental Research and Practical Applications. *J. Mater. Chem.* **2009**, *19* (47), 8907−8919.

(17) Choudhary, K.; Zhang, Q.; Reid, A. C. E.; Chowdhury, S.; Van Nguyen, N.; Trautt, Z.; Newrock, M. W.; Congo, F. Y.; Tavazza, F. Computational Screening of High-Performance Optoelectronic Materials Using OptB88vdW and TB-MBJ Formalisms. *Sci. Data* **2018**, *5*, 180082.

(18) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (Oct), 2825−2830.

(19) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including Crystal Structure Attributes in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96* (2), 24104.

(20) Wooten, F. *Optical Properties of Solids*; Academic press, 2013.

(21) Tran, F.; Blaha, P. Importance of the Kinetic Energy Density for Band Gap Calculations in Solids with Density Functional Theory. *J. Phys. Chem. A* **2017**, *121* (17), 3318−3325.

(22) Guan, S.; Yang, S. A.; Zhu, L.; Hu, J.; Yao, Y. Electronic, Dielectric and Plasmonic Properties of Two-Dimensional Electride Materials X2N (X = Ca, Sr): A First-Principles Study. *Sci. Rep.* **2015**, *5* (1), 12285.

(23) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (Nov), 2579−2605.