ACS APPLIED MATERIALS & INTERFACES

www.acsami.org

Research Article

Active Discovery of Donor:Acceptor Combinations For Efficient Organic Solar Cells

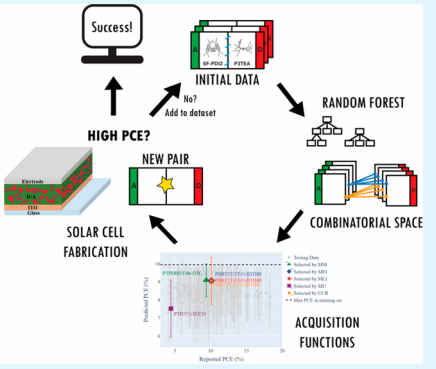
Prateek Malhotra, Juan C. Verduzco, Subhayan Biswas, and Ganesh D. Sharma*

Cite This: https://doi.org/10.1021/acsami.2c18540

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The structural flexibility of organic semiconductors offers vast a search space, and many potential candidates (donor and acceptor) for organic solar cells (OSCs) are yet to be discovered. Machine learning is extensively used for material discovery but performs poorly on extrapolation tasks with small training data sets. Active learning techniques can guide experimentalists to extrapolate and find the most promising D:A combination in a significantly small number of experiments. This study uses an active learning technique with a predictive random forest model to iteratively find the most optimal D:A combinations in the search space using various acquisition functions. Active learning results with five different acquisition functions (MM, MEI, MLI, MU, and UCB) are compared. Results reveal that acquisition functions that combine exploitation and exploration (MEI, MLI, and UCB) perform far better than purely exploiting (MM) and purely exploring (MU) acquisition functions. Interestingly, the proposed model can overcome the bottleneck of extrapolating small training data sets and find most promising D:A combinations in relatively fewer experiments.



KEYWORDS: *organic solar cells, power conversion efficiency, donor:acceptor combinations, machine learning, active learning, acquisition function*

1. INTRODUCTION

Organic solar cells (OSCs) have shown remarkable progress in the past decade and have gained much attention for being lightweight, flexible, transparent, and low-cost alternatives to conventional solar cell technology.[1−6] Power conversion efficiency (PCE) in the range of 18−19% for bulk heterojunction (BHJ) based OSCs has already been achieved[7−10] with the emergence of nonfullerene small molecule acceptor,[11,12] particularly Y-series small molecules.[13] By selecting suitable donors and acceptors with complementary absorption and matching frontier molecular orbitals (FMOs), the PCE of OSCs can be enhanced by as much as 20%.[2,14,15]

Nowadays, machine learning (ML) is gaining much attention, given its ability to accelerate productivity and material discovery.[16,17] ML models have been used in OSCs to investigate novel active materials and the information concealed in their chemical structures.[18] Traditionally, trial-and-error approaches based on intuition have remained the primary way to design novel materials. It would take years to explore a large chemical space of materials. However, with the evolution of ML approaches, scientists can now explore chemical space and its properties much more efficiently in terms of time and money. Interest in ML is increasing in material science-related fields because of the availability of massive data sets, improved algorithms, and exponentially increasing computing power. Various ML models are used for predicting power conversion efficiency (PCE),[19−36] short circuit current density (J_{SC}),[19,22,33,35] open-circuit voltage (V_{OC}),[19,22,33,35,37] fill factor (FF)[19,35] nonradiative voltage loss (ΔV_{NR}),[38] and frontier molecular orbitals (FMO).[13,22] Studies on high-throughput screening by creating a large search test space have been performed to discover potential OSC candidates.[36,39,40]

Organic semiconducting materials have tremendous scope for structural flexibility and rich design space, allowing a wide range of optoelectronic characteristics and FMOs to be tuned.[41] Innumerable OSC materials can be synthesized from derivatives of commonly used donor and acceptor materials by altering their donor moiety, acceptor moiety, side chain, core, and end-capping group.[42] This leads to many possible donor:acceptor (D:A) combinations for OSCs, and studies have been performed earlier using ML on systematically created large search spaces.[36,39,43] ML algorithms learn structure−property relationships[20,26] from training sets and offer an accelerated method for virtual screening of materials. A reliable supervised ML model, like regression, frequently needs

Received: October 21, 2022
Accepted: November 18, 2022

a huge amount of training data, typically obtained from expensive simulations or many experiments. If the available training data set is large enough, then feature engineering is a way to improve extrapolation capability of ML model. ML models usually struggle with extrapolation tasks, and training methods like leave-one-group-out (LOGO) cross-validation have been proposed to improve extrapolation capability.[44] However, extrapolation could be a problem with small training data sets, as it is challenging to generate a training set that precisely captures the chemical space of the large search space.[44] Provided with a large data set for training, variational autoencoders (VAE)[45] can be used to generate new active materials for creating a large search space.

Moreover, only predictions do not make clear sense after training the model on a small data set; associated uncertainties are also required to get confidence in the associated predictions.[46] In this study, a distinct technique is adopted that delivers accurate prediction and advises the selection of the ideal candidate data points to test next. This strategy is called active learning or sequential learning. Recently active learning has received much attention in a wide range of applications such as alloys,[47] thermoelectrics,[48] batteries,[49,50] OLEDs,[51] and OSCs[52] other organic semiconducting materials.[53]

Active learning builds a self-improving cycle that dynamically evaluates fresh data in order to maximize its predictive value. A major advantage of active learning over traditional ML algorithms is that it creates accurate models that query candidates to optimize the experimental design rather than relying on predictions for all values. Therefore, active learning is most beneficial in situations when data are few or challenging to gather, which is frequently the case in materials research. In fact, recent literature from a variety of areas has highlighted successful examples of active learning in discovery of new materials.[17,51,53] An important thing to note is that the buzzword "big data" refers to the search space, while the data set available for training is usually small and sparse, like in all other scientific domains. Thus, extrapolation is required, and traditional ML techniques might struggle to achieve it.

For the virtual screening using ML methods, analysis based on feature importance can help in identifying important structural fragments for achieving high PCE. A large search space is then constructed using these important structural fragments, and then predictions are made on this search space using the original model.[36,54,55] Hence, all the structural fragments involved in the search space are already available in the training set, and virtual screening can be done effectively. In our model, the training set consists of only a few materials and is familiar with only a few structural fragments. Thus, an effective active learning technique is required to find out the most promising candidates by performing experiments iteratively.

Along with the immensely wide chemical space of organic semiconductors, many parameters need to be tuned while considering any D:A combination. Some examples of tuning parameters include the following: (1) proper alignment of FMOs,[56] (2) complementary absorption spectra of donor and acceptors,[57] (3) miscibility of donor and acceptor,[58] (4) ratio of donor and acceptor in the BHJ active layer,[59] (5) thickness of electron transport layer (ETL), hole transport layer (HTL), and BHJ active layer,[60] and (6) annealing temperatures of ETL, HTL, and BHJ active layer.[61] Therefore, optimization of OSC with any specific D:A combination requires a lot of effort. It is not experimentally feasible to explore all possible D:A combinations manually in the lab since the process is expensive and time-consuming.[27] Thus, active learning techniques can guide experimentalists iteratively toward the most promising D:A combinations. The active learning model trained on the initial training data set is used to get predictions and sample-wise uncertainty estimates of the search space. Based on these predictions and uncertainty estimates of the search space, information acquisition functions select the most optimal candidate. Acquisition functions select the candidates by balancing exploitation and exploration. Exploration focuses on areas with high uncertainty, while exploitation emphasizes scenarios where the objective function is predicted to be maximized.[48] After each cycle, candidates selected by the acquisition function are included in the training set, and a new cycle is initiated in search of better candidates. With each successive iteration, knowledge gained by the model increases, prediction error decreases, and the model gets closer to maximizing the target property. The active learning approach used in this study is based on random Forest with Uncertainty Estimates for Learning Sequentially (FUELS) framework[48] by Ling et al. Such a framework does not require dimensionality reduction and is suitable for high-dimensional input parameter space. On the other hand, an active learning technique such as Bayesian optimization struggles with high dimensional data and requires dimensionality reduction. In this study, after dimensionality reduction using principal component analysis (PCA), results reveal that 60 PCA components are required to explain the complete variance as shown in Figure S1. Thus, Bayesian optimization might not be as effective.[48]

This work builds an active learning model that starts with an initial training set of D:A combinations and information acquisition functions to query the most optimal D:A combination from the search space. A data set of 200 unique D:A combinations is manually selected from the literature. All the descriptors for donors and acceptors for our training set are calculated using the RDKit python package.[62] In this work, we compare five acquisition functions: (1) maximum mean (MM), (2) maximum expected improvement (MEI), (3) maximum likelihood improvement (MLI), (4) maximum uncertainty (MU), and (5) upper confidence bound (UCB). To study the effect of dimensionality reduction, a similar pipeline is used for 60 pincipal components (PCs) and 40 PCs training sets. Since the data set is small, different D:A combinations explored by the acquisition functions are easily visualizable using t-distributed stochastic neighbor embedding (t-SNE) plot. The potential of active learning model is also examined on a newly published data set[63] of 1318 unique D:A combinations. This work aims to find the most promising D:A combination for achieving high PCE in a significantly small number of experiments.

2. EXPERIMENTAL SECTION

2.1. Data Gathering. A data set of 200 unique D:A combination is manually collected from the literature where all the donors are polymer, and all the acceptors are nonfullerene small molecule acceptors (NFSMAs). The number of unique donors is 70, and the number of unique acceptors is 95. The complete data set is provided in Table S1. Figure 1 represents the distribution of PCE in our data set.

To create a training set for active learning model, chemical structures of all donors and acceptors were drawn on ChemDraw software to retrieve their SMILES (simplified molecular-input line-entry system) strings.[64] SMILES strings define a chemical structure in

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

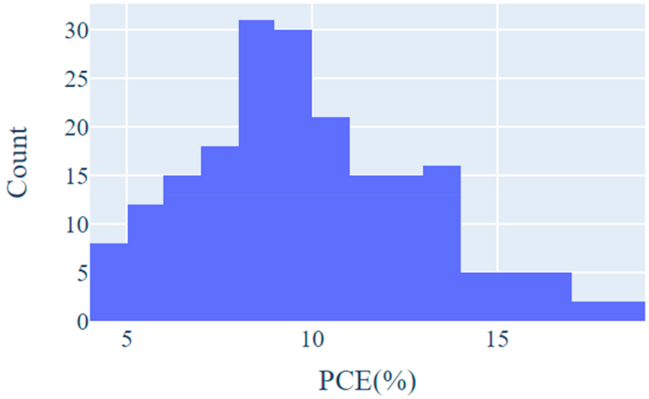


Figure 1. Distribution of PCE in manually collected data set of 200 unique donor:acceptor combinations.

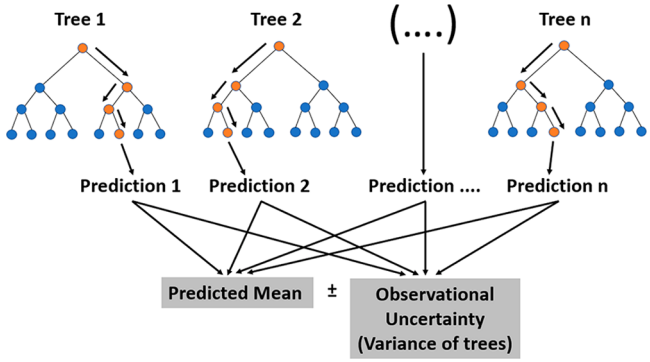


Figure 2. Calculation of predicted mean and sample-wise uncertainty estimates (observation uncertainty or variance of trees) by random forest regressor (Lolopy version 1.2.0). Predicted mean is the average of prediction by all trees, and observational uncertainty is the variance of prediction by all the trees.

a machine-readable format (ASCII strings). To express some of the information embedded in the chemical structures of donor and acceptor materials in a machine-readable format, a diverse set of molecular descriptors is required. Using SMILES strings, 208 molecular descriptors were calculated for each donor and acceptor using RDKit python package.[62] The RDKit package calculates constitutional, connectivity, MOE-type, molecular property, and topological descriptors.[65] Examples of commonly used descriptors are molecular weight, number of aliphatic or aromatic rings, number of rotatable bonds, fraction of sp^3 hybridized carbon atoms, and so on.

An abundance of associated features in the generated data set might harm the model's efficiency and accuracy. In order to avoid the curse of dimensionality, feature selection is required to simplify models and increase their interpretability as well as their training efficiency. This study removed unnecessary features from the training set by dropping constant features and features with high correlation coefficients.

2.2. Random Forest and Uncertainty Estimates. A random forest model is a supervised learning algorithm in which a collection of decision trees are generated from randomly selected subsets of rows and descriptors. This process of random selection with replacement is called "Bagging" or "Bootstrap Aggregation". Whenever a decision tree is created to its complete depth, it leads to overfitting (high variance). However, multiple decision trees are combined in a random forest, and variance gets reduced. Prediction of the model is given by the mean of output from all decision trees.

Ling et al. used a random Forest with Uncertainty Estimates for Learning Sequentially (FUELS) framework[48] for uncertainty quantification and showed its potential as a tool to discover new materials. Based on their work, through the python library Lolopy version (1.2.0), we were able to get predicted mean along with sample-wise uncertainty estimates by evaluating observational variance in the trees of the forest. An illustration for calculating predicted mean and observational uncertainty is shown in Figure 2.

These sample-wise uncertainty estimates should be well-calibrated, and a calibration check can be done by using normalized residual. Normalized residual (r_n) is defined as the difference between predicted value $\{\hat{f}(x)\}$ and actual value$\{f(x)\}$ divided by the observational uncertainty $\{\sigma^2(x)\}$. $r_n = \frac{\hat{f}(x) - f(x)}{\sigma^2(x)}$

In our random forest model, the number of estimators (trees) was set to 350, and all the trees were created to their complete depth so that our model could capture most of the information from training data. To visualize the uncertainty estimates calculated by the random forest regressor by Lolopy, we splitted the data set into 90:10 train/test ratio. Results are shown in Figure 3a, where blue points represent training data, and red points represent predicted mean of test data along with their observational uncertainty. Mean absolute error (MAE) for the test set came out to be 1.67%. To check the uncertainty estimate accuracy, 10-fold cross-validation is performed on the complete data set to create a distribution of normalized residual. Perfect calibration of uncertainty estimate would result in the distribution of normalized residual to be a Gaussian with zero mean and unit standard deviation. Our distribution for normalized residual is shown in Figure 3b, which is roughly Gaussian with 0.0231 mean and 1.087 standard deviation.

2.3. Acquisition Functions. Acquisition functions are the mathematical expressions that are used to select candidates based on predicted mean and sample-wise uncertainty estimates of the search space. In this work, predicted mean and uncertainty estimates are calculated using the Random Forest regressor by Lolopy version 1.2.0. Acquisition functions may use exploitation, exploration, or a combination of both, depending on the use case. Exploiting functions select the most similar candidates irrespective of their uncertainties, and exploring functions choose the candidates with high uncertainty to gain more knowledge. Based on the predicted PCE and their associated uncertainties by random forest model, the acquisition function will decide which D:A combination to reveal next. The idea is to keep on iterating until the best candidate is found. To get a better candidate in very few iterations (experiments), candidate with the highest predicted mean and lowest uncertainty should be chosen. However, if budget allows performing a number of iterations (experiments), then the candidate with the highest uncertainty should be chosen because it will provide the model with new insights. Five established acquisition functions from the literature are used, namely, maximum mean (MM), maximum expected improvement (MEI), maximum likelihood improvement (MLI), maximum uncertainty (MU), and upper confidence bound (UCB).

2.3.1. Maximum Mean (MM). This acquisition function simply selects the candidate with the highest predicted mean value out of all available candidates in the test data set. This function is greedy and looks for similar candidates that may enhance the target variable. This can be referred to as exploitation.

MM:

$$x^* = \text{argmax}\, E[M(x_i)] \quad (1)$$

where x^* represents the selected candidate by the acquisition function, x_i represents the complete search space, $E[M(x_i)]$ is the mean prediction of the model at point x_i, and the "argmax" function returns the index of maximum value along an axis.

2.3.2. Maximum Expected Improvement (MEI). MEI combines exploitation and exploration to find the most optimal candidate. This function uses the mean prediction and uncertainty estimates to draw the probability distribution function (PDF) and cumulative distribution function (CDF) at each test point to select the optimal candidate.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

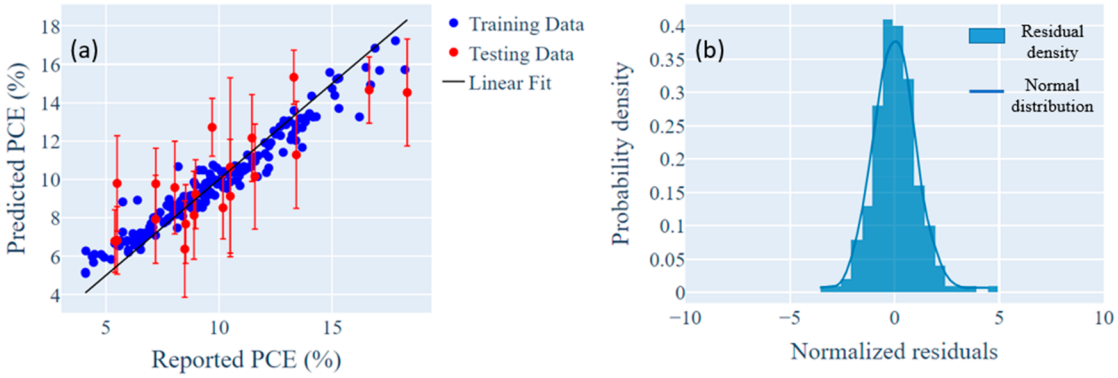


Figure 3. (a) Visualization of uncertainty estimates in predicting PCE. Blue points represent training data; red points represent the predicted mean of test data and their observational uncertainty. The black line represents linear fit. (b) Probability density of normalized residual to check uncertainty estimate accuracy.

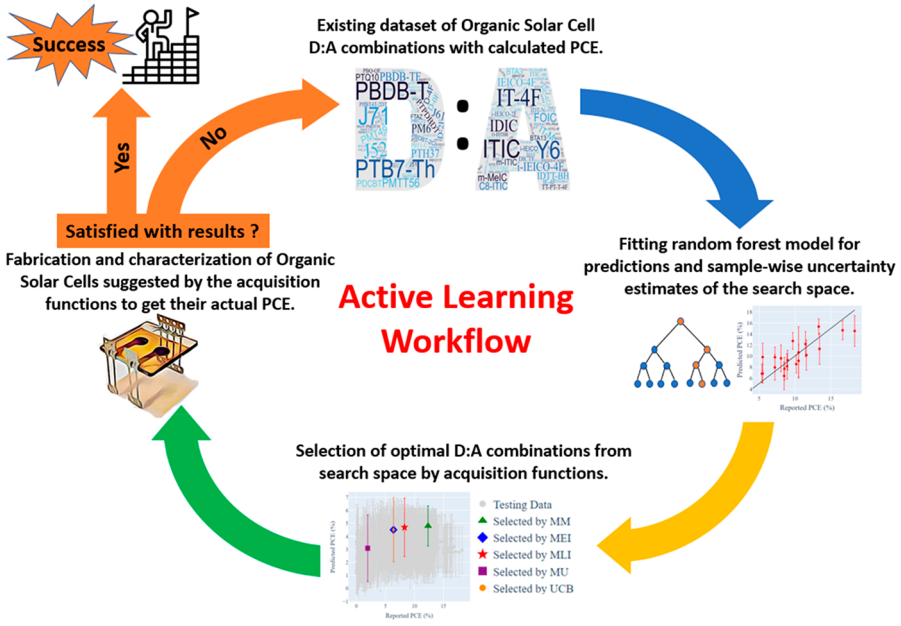


Figure 4. Schematic representation of active learning workflow. Random forest model is trained on the initial data set of D:A combinations to make predictions on the search space along with sample-wise uncertainty estimates. The acquisition function selects the most optimal candidate for higher PCE possibility based on mean predictions and sample-wise uncertainty estimates. Selected candidates are then fabricated in the lab and characterized to get actual PCE. If the user is not satisfied with the result, then the selected candidate is added to the training set, and the whole process is reinitiated. If the user is satisfied with the result, then this is considered a success.

MEI:

$$x^* = \text{argmax}\, \rho(E[M(x_i)] - E[M(x_{\text{best}})],\ \sigma^2[M(x_i)])$$

$$\text{with}\quad \rho(z, s) = \begin{cases} s\Phi'\left(\frac{z}{s}\right) + z\Phi\left(\frac{z}{s}\right) & s > 0 \\ \max(z, 0) & s = 0 \end{cases} \tag{2}$$

where, x_{best} is the current best candidate in the data set, σ^2 is the uncertainty (variance of the estimators), Φ' is PDF, and Φ is CDF.[66,67]

2.3.3. Maximum Likelihood Improvement (MLI). Based on uncertainty, this function selects the candidate most likely to give a target value better than the best candidate in the available train data set. This function also gives combined attention to exploitation and exploration.[68]

MLI:

$$x^* = \text{argmax}\, \frac{E[M(x_i)] - E[M(x_{\text{best}})]}{\sigma^2[M(x_i)]} \tag{3}$$

2.3.4. Maximum Uncertainty (MU). MU selects the candidate with the highest uncertainty, independent of their predicted mean. This can lead to selection of candidates that are not optimal for study but are optimal for capturing the search space.

MU:

$$x^* = \text{argmax}\, \sigma^2[M(x_i)] \tag{4}$$

2.3.5. Upper Confidence Bound (UCB). UCB selects the candidate based on the maximum of predicted mean plus associated uncertainty with a tuning parameter (K).[68] For this study, we have selected tuning factor = 1.

UCB:

$$x^* = \text{argmax}\, (E[M(x_i)] + K*\sigma^2[M(x_i)]) \tag{5}$$

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

These five acquisition functions are compared to determine which one performs best, i.e., finds the best candidate faster.

2.4. Active Learning Workflow. A lot of studies have been done to predict PCE of organic solar cells by ML using the chemical structure of donors and acceptors as descriptors.[16,17,63] The most straightforward ML approach is to create a small data set of D:A combinations and their reported PCE and train a regression model to predict PCE of unexplored D:A combinations, but the potential problem with this approach is extrapolation since it is very challenging to create a training set that can accurately capture the chemical space of the entire search space (testing set). To extrapolate means to find a candidate that is better than anything available in the training data set.

Machine learning is most suitable for large data sets, but scientific data sets are usually small and sparse. In this work, we have chosen a different approach that not only makes accurate prediction but also guide the selection of the most optimal training data points. The size of the initial training data set is only bound to the creation of a model that can capture some of the relationships between our descriptors and PCE. Figure 4 represents the active learning workflow for finding the most promising D:A combination for high PCE. The workflow starts with fitting a random forest model on a small existing D:A combination data set with known PCE. The fitted model is then used to get the search space's mean predictions and sample-wise uncertainty estimates. With these values, the acquisition function identifies the most optimal D:A combination to achieve higher PCE. The D:A combinations selected by the acquisition function are fabricated in the lab and characterized to get actual PCE. If the user is not satisfied with the results, then the selected candidate is added to the training set, and the cycle is reinitiated. If the results are satisfactory, then it is considered a success. Data increase with each successive iteration (experiment). Hence, predictions will become more accurate, and the research goal can be achieved much faster. Aiming to achieve this goal, we want to show that we can dramatically reduce the number of iterations (experiments) to reach our design goal by using active learning techniques.

3. RESULT AND DISCUSSION

Since this work focuses on D:A combinations, an initial training D:A combination set should be selected such that none of their donor or acceptor is involved in a PCE of greater than 10%. If this step is bypassed, then pure exploiting acquisition functions such as MM would have an edge in finding the best candidate in very few iterations.

To demonstrate active learning, random forest model is trained on 10 randomly selected D:A combinations are given in Table 1 to predict PCE along with sample-wise uncertainty estimates for the test set (190 D:A combination). Based on predicted PCE and sample-wise uncertainty estimates, the acquisition function will decide which D:A combination to reveal next. Figure 5 represents candidates selected by all five acquisition functions in the first iteration of active learning to find the candidate who is most likely to give a higher PCE. Search space is represented in gray color, candidates selected by the acquisition function are represented in different colors, and maximum PCE in the training set is represented by black dotted line (9.95%). Since random forest cannot extrapolate, all the predicted mean values will lie between upper and lower bounds of the initial training data set.

Table 1. Randomly Selected 10 D:A Combinations with PCE Less Than 10%

s. no.	donor	acceptor	PCE (%)
1	PffBT4T-2DT	FBR	7.80
2	PffBT4T-2DT	IDTBR	9.95
3	P3TEA	SF-PDI2	9.50
4	BDT-ffBX-DT	SFPDI	6.20
5	PMOT39	i-IEICO-2F	6.00
6	PTPDBDT	F-ITIC	8.80
7	PTPDBDT	Cl-ITIC	9.50
8	PffBT-T3	TPPz-PDI4	6.90
9	PDBT-T1	IDIC	9.20
10	PTFBDT-BZS	IDIC	8.06

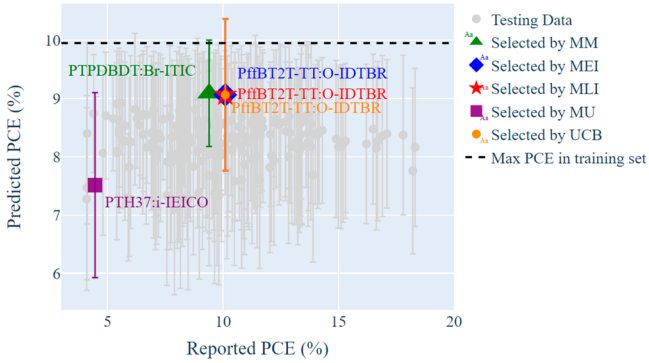


Figure 5. Illustration of first iteration results. Predictions of random forest model trained on randomly selected data (Table 1) along with sample-wise uncertainty estimates are shown in gray color. In the first iteration, candidates selected by MM, MEI, MLI, MU, and UCB acquisition function are colored, and the error bar represents observational uncertainty. Black dotted line represents the maximum PCE in the training set (9.95%). Acquisition functions that combine both exploration and exploitation (MEI, MLI, and UCB) outperform pure exploitation (MM) and pure exploration (MU). Interestingly, MEI, MLI, and UCB selected the same candidate.

As seen in Figure 5, MM (pure exploitation) selected the PTPDBDT:Br-ITIC combination with predicted mean PCE (9.08%) and observational uncertainty (±0.91%). MM chose this combination because of the highest predicted mean, completely ignoring uncertainty. Since all the PCEs in the training set are less than 10%, due to bootstrap aggregation, none of the mean predictions can exceed 10% PCE. MU (pure exploration) selected PTH37:i-IEICO combination with predicted mean PCE (7.5%) and observational uncertainty (±1.58%). MU chose this combination because of the highest uncertainty and completely ignored the predicted mean. In D:A combination chosen by MU, the difference between reported PCE and predicted PCE is high, and the associated uncertainty is also high. This again confirms good calibration of uncertainty estimates. Surprisingly, acquisition functions that combine both exploration and exploitation (MEI, MLI, and UCB) outperform in the very first iteration and are able to find out the candidate better than anything in the training set (uncertainty is surpassing the dotted line (9.95%). MEI, MLI, and UCB selected the same candidate in the first iteration (PffBT2T-TT:O-IDTBR) with predicted mean PCE (9.06%) and observational uncertainty (±1.3%). It is important to note that general practice is to fabricate the cell suggested by the acquisition function, but instead of going to the lab, we will reveal the PCE as reported in the literature. If the results are not satisfactory, then a second iteration is initiated. For running the second iteration, the candidate selected by the acquisition function in the first iteration is added to the training data set. Now the training set has 11 D:A combinations. Again, the model is fitted to predict the remaining test set (189 D:A combination), and the acquisition function selects the next suitable candidate. This loop goes on for n number of iterations until the best candidate is found.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

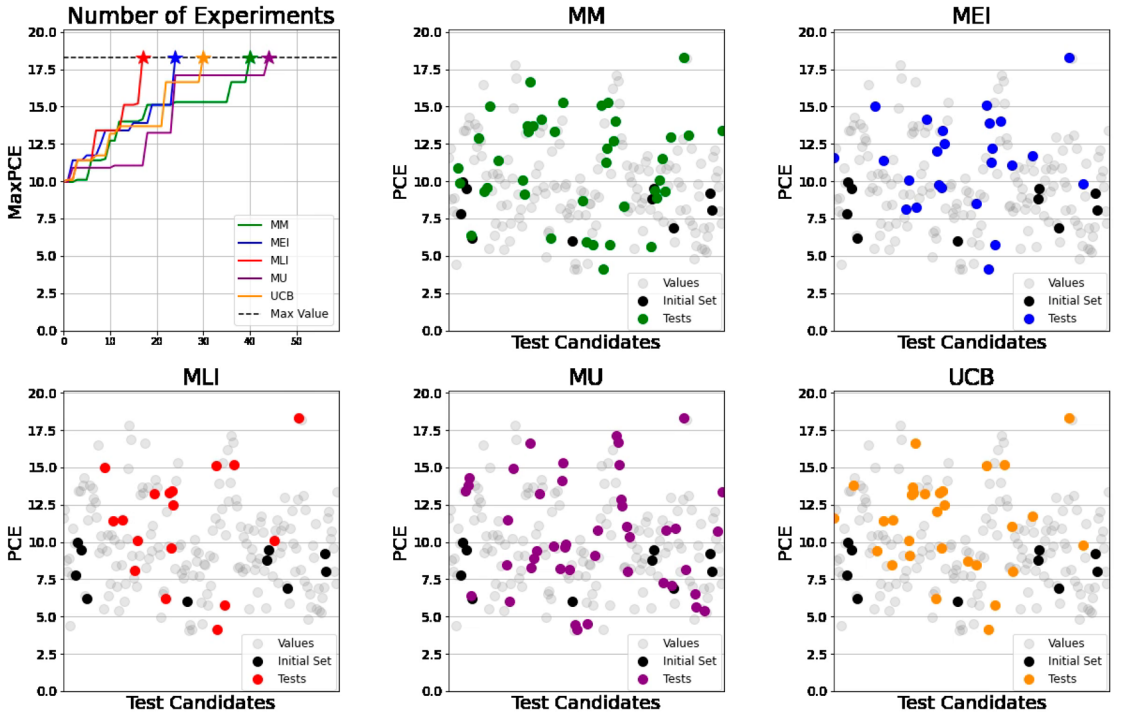


Figure 6. Comparison of different acquisition functions. The initial training data set is common for all acquisition functions and is represented by black dots. Gray dots represent search space, and colored dots represent candidates that have been identified by the acquisition function in the search space. The first panel in the top left depicts a trend line that tracks the optimal candidate for individual acquisition function with increasing experiments. Dotted line represents the highest PCE in the study search space (18.32%). When the acquisition function finds the candidate with highest PCE, the trend line stops with a star and stops discovering further.

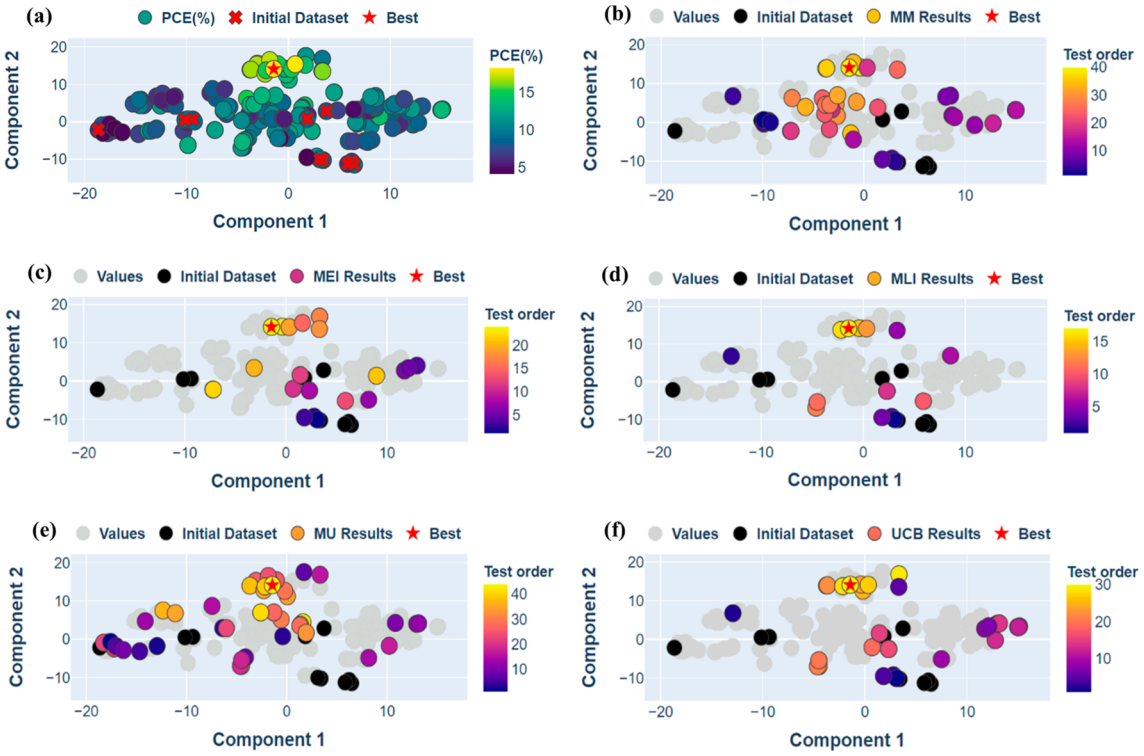


Figure 7. (a) t-SNE projection of 200 unique D:A combination having wide structural variety; color bar represents PCE (%). Cross marks represent training data set, and star mark represents top candidate with PCE 18.32%. (b–f) The same t-SNE plot for different acquisition functions; color bar represents the test order by which optimal candidates were identified by the corresponding acquisition function. Gray dots represent the search space. Black dots represent the initial training data set, and the star mark represents the top candidate. Number of iterations by each acquisition function are as follows: MM (40), MEI (24), MLI (17), MU (44), and UCB (30).

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

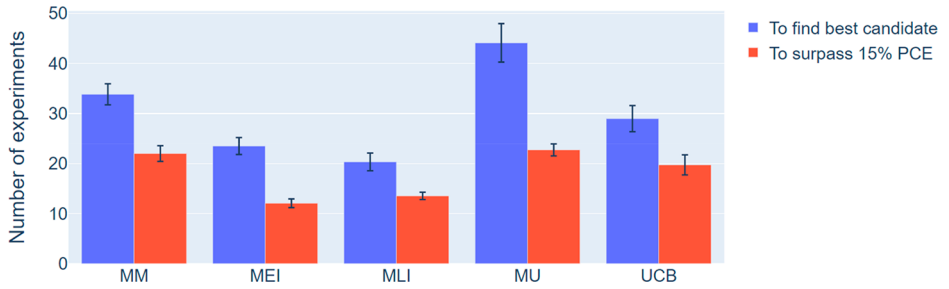


Figure 8. Blue bars represent the average number of experiments carried out by each acquisition function to discover the best candidate in the search space, while the red bars represent the average number of experiments carried out by each acquisition function to surpass 15% PCE threshold. Error bars represent standard error $E(x) = \frac{\sigma}{\sqrt{30}}$. MLI performed best, followed by MEI, while MM and MU performed worst for our study.

Results for each acquisition function are shown in Figure 6. Black dots represent the initial training data set, and it is same for all acquisition functions. Gray dots represent the search space, and colored dots represent candidates identified by different acquisition functions in the search space. The top left panel depicts a trend line that tracks the optimal candidate for individual acquisition function with an increasing number of experiments. When the acquisition function finds the candidate with highest PCE, the trend line stops with a star and does not discover further.

Results show that the acquisition functions that combine exploitation and exploration (UCB, MLI, and MEI) consistently outperform the pure exploiting (MM) and pure exploring (MU) acquisition functions. The trajectory of all acquisition functions for the results shown in Figure 6.

The in-depth optimization process described by the trend line in the first panel of Figure 6 can be visualized using t-SNE plot for D:A combination. t-SNE takes a high-dimensional data set and reduces it to a two-dimensional graph that preserves a lot of original information. The two dimensions of t-SNE plot (components 1 and 2) reflect the difference in the candidates in the feature space. Molecules with similar chemical structures will be clustered together in this simplified two-dimensional graph, while those with distinct chemical structures will spread apart. In Figure 7, t-SNE plot is created for 200 unique D:A combinations using morgan fingerprints by RDKit python package. Morgan fingerprints[69] (nbits = 2048 and radius = 2) were calculated separately for donors and acceptors and concatenated.

Figure 7a represents the t-SNE projection of 200 unique D:A combination having wide structural variety, the color bar represents PCE (%), cross marks represent the training data set, and the star mark represents the best candidate with PCE of 18.32%. Figure 7b−e represents the same t-SNE plot for different acquisition functions, and the color bar represents the test order by which optimal candidates were identified by the corresponding acquisition function described in trend lines of Figure 6. Figure 7b−e shows that gray dots represent the search space, black dots represent the initial training data set, and star marks represent the best candidate. Black dots representing the initial training data set are spread apart, indicating distinct chemical structures of donors and acceptors in training set and are mentioned in Table S2 along with chemical structures of donors and acceptors. For the MM function in Figure 7b, each successive iteration exploits the training set and identifies the chemically similar candidates for the next iteration. MU function in Figure 7e tries to explore the chemical space as much as possible to include high uncertainty candidates so that model can learn much information. With color bar in Figure 7e, exploration is also clearly visible with test order. Both MM and MU functions cover almost the whole structural space of D:A combinations concerning exploitation and exploration indicated by their test orders. MEI, MLI, and UCB functions in Figure 7c,d,f combine exploitation and exploration to find the most optimal candidate in the search space to reach the top candidate in the least number of iterations. Thus, these three functions target the specific regions of the chemical space where relatively higher predictions of PCE are possible and select the most optimal candidates to reach the top. Best results are obtained by MLI in Figure 7d. For Figure 7b−e, D:A combinations associated with the test order are mentioned in Tables S3−S7. Number of iterations by each acquisition function to find the best candidate are as follows: MM (40), MEI (24), MLI (17), MU (44), and UCB (30).

In order to quantify the relative performance of acquisition functions, we repeated this experiment 30 times with an initial set of 10 randomly selected points. The results are shown in Figure 8 below. Since these acquisition functions can easily get trapped into local maxima (especially MM and MU) for a considerable number of iterations, it is also important to note how many iterations are required to surpass 15% PCE threshold. The blue bars represent the average number of experiments carried out by each acquisition function in order to discover the best candidate in the search space, while the red bars represent the average number of experiments carried out by each acquisition function in order to surpass 15% PCE threshold. Error bars represent standard error $E(x) = \frac{\sigma}{\sqrt{30}}$. MLI performed best, followed by MEI, while MM and MU performed worst for our study to find the best candidate. MEI performed best to surpass the 15% PCE threshold, followed by MLI. These results are also tabulated in Table 2.

The model's performance is also studied for different descriptors sets (RDKit, Mordred,[70] and mix of both), and the effect of dimensionality reduction is also studied by using

Table 2. Average Number of Experiments and Standard Error over 30 Experiments for All Acquisition Functions

	MM	MEI	MLI	MU	UCB
to find best candidate	34 ± 2	24 ± 2	20 ± 2	44 ± 4	29 ± 3
to surpass 15% PCE	22 ± 2	12 ± 1	14 ± 1	23 ± 1	20 ± 2

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

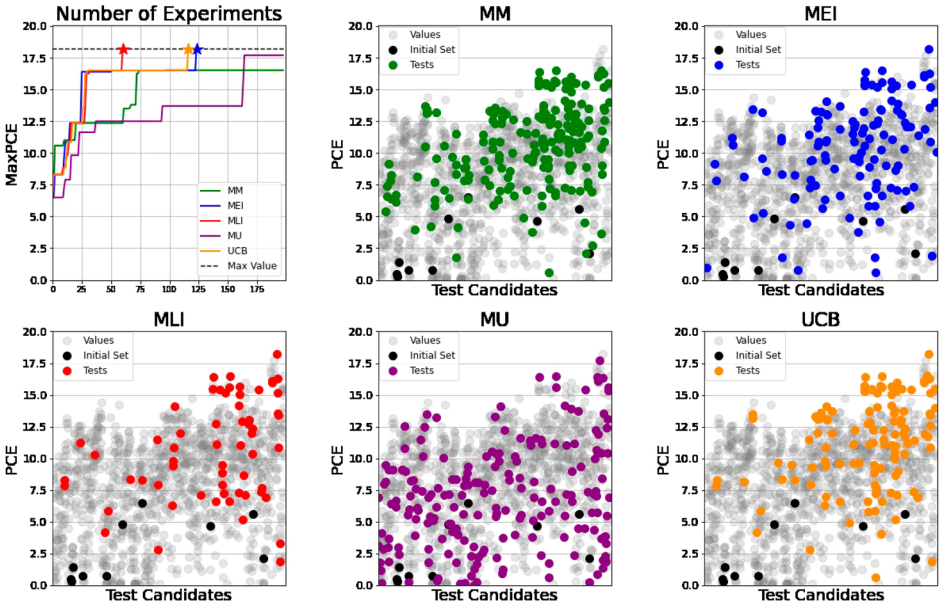


Figure 9. Comparison of different acquisition functions for 1318 data sets. Initial training data sets are common for all acquisition functions and are represented by black dots. Gray dots represent search space, and colored dots represent candidates that have been identified by the acquisition function in the search space. The first panel depicts a trend line that tracks the optimal candidate for individual acquisition function with an increasing number of experiments. When the acquisition function finds the candidate with the highest PCE, trend line stops with a star and stops discovering further.

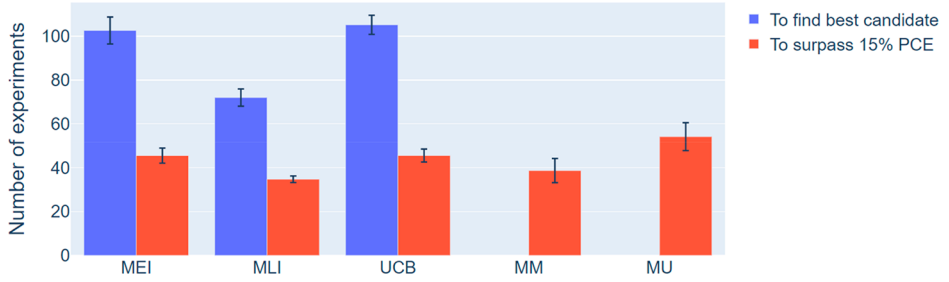


Figure 10. Blue bars represent the average number of experiments carried out by each acquisition function to discover the best candidate in the search space, while the red bars represent the average number of experiments carried out by each acquisition function to surpass a 15% PCE threshold. Since MM and MU cannot find the best candidate within 200 iteration budget, their results are not displayed. Error bars represent standard error $E(x) = \frac{\sigma}{\sqrt{30}}$. MLI performed best, followed by MEI, while MM and MU performed worst in our study.

60 and 40 PCs. Results are shown in Figures S2−S4. Rdkit descriptors perform comparatively well, and the model is computationally less expensive because of fewer descriptors. RDKit descriptors and mix descriptors performed well without dimensionality reduction by using MEI, MLI, and UCB acquisition functions. For Mordred descriptors, dimensionality reductions gave better results. Moreover, we have compared our results with Bayesian optimization (using Gaussian process regressor as surrogate model with expected improvement (EI) acquisition function) and found that random forest as a surrogate model performs better for our study. Results for GPR as surrogate model are shown in Figure S5. On the other hand, GPR with EI acquisition function gave highly mixed results, and the best results were achieved by mix of RDKIT and Mordred descriptors.

To prove the potential of our model, we applied an end-to-end model to a newly published data set with 1318 unique D:A combinations,[63] and their distribution of PCE is shown in Figure S6. We applied the same strategy for selecting the training set (randomly choosing 10 D:A combinations such that none of the donors or acceptors are involved in a PCE of greater than 10%). Using the same restrictions as earlier initial training set is chosen randomly and is shown in Table S8. In randomly selected initial data set for training, 6 out of 10 values have PCE smaller than 2%, making it a very rough training set to check the potential of our active learning model. Figure S7 represents first iteration selections by the acquisition functions. In this study as well, all the predicted mean values in the first iteration will lie between upper and lower bound of initial training set, and cannot exceed 6.5% (highest PCE for the training set). Acquisition functions that combine exploitation and exploration (MEI, MLI, and UCB) outperform pure exploitation (MM) and pure exploration (MU). MEI, MLI, and UCB select the candidate whose uncertainty shows the potential to surpass the PCE of the current best candidate in the training set.

Results for each acquisition function with a trend line are shown in Figure 9. Number of iterations by each acquisition function to find the best candidate are as follows: MEI (123), MLI (60), and UCB (116), while MM and MU were not able to find the best candidate within the 200 iteration budget.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

Since the acquisition function is more likely to get trapped in local maxima for many iterations in a larger data set, we have compared iterations to get the best candidate with iterations to surpass the 15% PCE mark. For the 1318 data sets, relative performance of acquisition function is quantified by repeating the experiment 30 times, and the results are shown in Figure 10 below. This time model ran for 200 iterations, and acquisition function that involves both exploitation and exploration (MEI, MLI, and UCB) outperforms solely exploiting (MM) and solely exploring (MU). Due to a bigger search space this time, MM and MU could not find the best candidate within the 200 iteration budget. This proves that a simple strategy of pure exploitation or pure exploration cannot find promising candidates within a few experiment budget for such a use case. These results are also tabulated in Table 3.

Table 3. Average Number of Experiments and Standard Error over 30 Experiments for All Acquisition Functions[a]

	MM	MEI	MLI	MU	UCB
to find best candidate		103 ± 6	72 ± 4		105 ± 4
to surpass 15% PCE	39 ± 6	46 ± 3	35 ± 2	54 ± 6	46 ± 3

[a]Since MM and MU could not find the best candidate within 200 iteration budget, they are left blank.

4. CONCLUSION

Active learning technique based on random forest with uncertainty estimates is used to find the most optimal D:A combination in the search space in the least possible iterations. This work demonstrates how active learning workflow can be used at the lab scale to find the most promising D:A combination in the fewest possible iterations (experiments). A data set of manually collected 200 unique D:A combinations is used along with their corresponding PCE. Since the whole purpose of active learning is to explore the vast search space, simply accessible descriptors (RDKit descriptors) are used. Active learning workflow starts with randomly chosen D:A combination such that none of the donors or acceptors is involved in PCE greater than 10%. Five acquisition functions (MM, MEI, MLI, MU, and UCB) are compared using a trend line for finding the best candidate in the search space. From the comparison of acquisition functions, results reveal that the best performing acquisition function combines both exploitation and exploration (MEI, MLI, and UCB), while purely exploiting (MM) and purely exploring (MU) functions perform worst. A t-SNE plot is used to represent how different acquisition function trend line traces the chemical search space of D:A combinations using exploitation and exploration. Performance of each acquisition function is quantified by running the experiment 30 times with the randomly selected initial training set, and the mean value is calculated along with the standard error. MLI (20 iterations) performs best, followed by MLI (24 iterations). The number of iterations taken by different acquisition functions to surpass 15% PCE threshold is also studied as acquisition functions are likely to get stuck into a local maxima for a considerable number of iterations. For surpassing the 15% PCE threshold, MEI performed best (12 iterations), followed by MLI (14 iterations). After getting satisfactory results, the model potential is examined with a recently published data set of 1318 unique D:A combinations, and comparable results were achieved. This proves the potential of our model for lab-based research with several tuning parameters. For applications of ML in OSCs, if the target variable is very complex and ML algorithms do not perform well on structural descriptors, then DFT calculated descriptors (FMO, band gap, etc.) can provide much more insight into ML algorithms and better results could be achieved. This active learning workflow can be used to iteratively guide experimentalists in finding most promising D:A combinations given a large search space.

■ ASSOCIATED CONTENT

Data Availability Statement

Data that support the finding of this study are available from the corresponding author upon reasonable request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsami.2c18540.

Data set of 200 unique D:A combinations with corresponding PCE manually collected from the literature; PCA analysis representing complete explanation of variance with 60 principal components (PCs); randomly selected 10 D:A combinations with PCE less than 10%; D:A combination selected by MM acquisition function, by MEI acquisition function, by MLI acquisition function, by MU acquisition function, and by UCB acquisition function, with corresponding test orders; randomly chosen initial training set with PCE less than 10% for the 1318 data set; number of experiments carried out to discover the best candidate in search space using random forest with RDKit descriptors, using random forest with Mordred descriptors, using random forest with mix of RDKit and Mordred descriptors, and using Gaussian process regressor (GPR) with RDKit descriptors, Mordred descriptors, and mix of both; PCE distribution for the 1318 unique D:A combinations and first iteration results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ganesh D. Sharma − *Department of Physics and Department of Electronics Engineering and Communication, The LNM Institute of Information Technology, Jaipur 302031 Rajasthan, India*; orcid.org/0000-0002-1717-0116; Email: gdsharma273@gmail.com, gdsharma@gmail.com

Authors

Prateek Malhotra − *Department of Physics, The LNM Institute of Information Technology, Jaipur 302031 Rajasthan, India*

Juan C. Verduzco − *School of Materials Engineering and Birck Nanotechnology Center, Purdue University, West Lafayette, Indiana 47907, United States*

Subhayan Biswas − *Department of Physics, The LNM Institute of Information Technology, Jaipur 302031 Rajasthan, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsami.2c18540

Author Contributions

All the authors contributed equally to this work.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are thankful to Prof. Alejandro Strachan, School of Materials Engineering, Purdue University, for his valuable suggestions. We acknowledge computational resources from nanoHUB and Purdue University through the Network for Computational Nanotechnology. Mr. Prateek Malhotra is thankful to the LNM Institute of Information technology for providing institutional fellowship.

■ REFERENCES

(1) Almora, O.; Baran, D.; Bazan, G. C.; Berger, C.; Cabrera, C. I.; Catchpole, K. R.; Erten-Ela, S.; Guo, F.; Hauch, J.; Ho-Baillie, A. W. Y.; Jacobsson, T. J.; Janssen, R. A. J.; Kirchartz, T.; Kopidakis, N.; Li, Y.; Loi, M. A.; Lunt, R. R.; Mathew, X.; McGehee, M. D.; Min, J.; Mitzi, D. B.; Nazeeruddin, M. K.; Nelson, J.; Nogueira, A. F.; Paetzold, U. W.; Park, N.; Rand, B. P.; Rau, U.; Snaith, H. J.; Unger, E.; Vaillant-Roca, L.; Yip, H.; Brabec, C. J. Device Performance of Emerging Photovoltaic Materials (Version 2). *Adv. Energy Mater.* **2021**, *11* (48), 2102526.

(2) Karki, A.; Gillett, A. J.; Friend, R. H.; Nguyen, T. The Path to 20% Power Conversion Efficiencies in Nonfullerene Acceptor Organic Solar Cells. *Adv. Energy Mater.* **2021**, *11* (15), 2003441.

(3) Duan, L.; Uddin, A. Progress in Stability of Organic Solar Cells. *Adv. Sci.* **2020**, *7* (11), 1903259.

(4) Hong, L.; Yao, H.; Cui, Y.; Ge, Z.; Hou, J. Recent Advances in High-Efficiency Organic Solar Cells Fabricated by Eco-Compatible Solvents at Relatively Large-Area Scale. *APL Mater.* **2020**, *8* (12), 120901.

(5) Wang, D.; Qin, R.; Zhou, G.; Li, X.; Xia, R.; Li, Y.; Zhan, L.; Zhu, H.; Lu, X.; Yip, H.; Chen, H.; Li, C. High-Performance Semitransparent Organic Solar Cells with Excellent Infrared Reflection and See-Through Functions. *Adv. Mater.* **2020**, *32* (32), 2001621.

(6) Wu, J.; Gao, M.; Chai, Y.; Liu, P.; Zhang, B.; Liu, J.; Ye, L. Towards a Bright Future: The Versatile Applications of Organic Solar Cells. *Mater. Reports Energy* **2021**, *1* (4), 100062.

(7) Li, C.; Zhou, J.; Song, J.; Xu, J.; Zhang, H.; Zhang, X.; Guo, J.; Zhu, L.; Wei, D.; Han, G.; Min, J.; Zhang, Y.; Xie, Z.; Yi, Y.; Yan, H.; Gao, F.; Liu, F.; Sun, Y. Non-Fullerene Acceptors with Branched Side Chains and Improved Molecular Packing to Exceed 18% Efficiency in Organic Solar Cells. *Nat. Energy* **2021**, *6* (6), 605−613.

(8) Liu, Q.; Jiang, Y.; Jin, K.; Qin, J.; Xu, J.; Li, W.; Xiong, J.; Liu, J.; Xiao, Z.; Sun, K.; Yang, S.; Zhang, X.; Ding, L. 18% Efficiency Organic Solar Cells. *Sci. Bull.* **2020**, *65* (4), 272−275.

(9) Cui, Y.; Xu, Y.; Yao, H.; Bi, P.; Hong, L.; Zhang, J.; Zu, Y.; Zhang, T.; Qin, J.; Ren, J.; et al. Single-Junction Organic Photovoltaic Cell with 19% Efficiency. *Adv. Mater.* **2021**, *33* (41), 2102420.

(10) Qin, J.; Yang, Q.; Oh, J.; Chen, S.; Odunmbaku, G. O.; Ouedraogo, N. A. N.; Yang, C.; Sun, K.; Lu, S. Volatile Solid Additive-Assisted Sequential Deposition Enables 18.42% Efficiency in Organic Solar Cells. *Adv. Sci.* **2022**, *9*, 2105347.

(11) Hou, J.; Inganas, O.; Friend, R. H.; Gao, F. Organic Solar Cells Based on Non-Fullerene Acceptors. *Nat. Mater.* **2018**, *17* (2), 119−128.

(12) Wadsworth, A.; Moser, M.; Marks, A.; Little, M. S.; Gasparini, N.; Brabec, C. J.; Baran, D.; McCulloch, I. Critical Review of the Molecular Design Progress in Non-Fullerene Electron Acceptors towards Commercially Viable Organic Solar Cells. *Chem. Soc. Rev.* **2019**, *48* (6), 1596−1625.

(13) Yuan, J.; Zhang, Y.; Zhou, L.; Zhang, G.; Yip, H. L.; Lau, T. K.; Lu, X.; Zhu, C.; Peng, H.; Johnson, P. A.; Leclerc, M.; Cao, Y.; Ulanski, J.; Li, Y.; Zou, Y. Single-Junction Organic Solar Cell with over 15% Efficiency Using Fused-Ring Acceptor with Electron-Deficient Core. *Joule* **2019**, *3* (4), 1140−1151.

(14) Upama, M. B.; Mahmud, M. A.; Conibeer, G.; Uddin, A. Trendsetters in High-Efficiency Organic Solar Cells: Toward 20% Power Conversion Efficiency. *Sol. RRL* **2020**, *4* (1), 1900342.

(15) Azzouzi, M.; Yan, J.; Kirchartz, T.; Liu, K.; Wang, J.; Wu, H.; Nelson, J. Nonradiative Energy Losses in Bulk-Heterojunction Organic Photovoltaics. *Phys. Rev. X* **2018**, *8* (3), 31055.

(16) Mahmood, A.; Wang, J.-L. Machine Learning for High Performance Organic Solar Cells: Current Scenario and Future Prospects. *Energy Environ. Sci.* **2021**, *14* (1), 90−105.

(17) Rodríguez-Martínez, X.; Pascual-San-José, E.; Campoy-Quiles, M. Accelerating Organic Solar Cell Material's Discovery: High-Throughput Screening and Big Data. *Energy Environ. Sci.* **2021**, *14* (6), 3301−3322.

(18) Yan, J.; Rodríguez-Martínez, X.; Pearce, D.; Douglas, H.; Bili, D.; Azzouzi, M.; Eisner, F.; Virbule, A.; Rezasoltani, E.; Belova, V.; Dörling, B.; Few, S.; Szumska, A. A.; Hou, X.; Zhang, G.; Yip, H.-L.; Campoy-Quiles, M.; Nelson, J. Identifying Structure−Absorption Relationships and Predicting Absorption Strength of Non-Fullerene Acceptors for Organic Photovoltaics. *Energy Environ. Sci.* **2022**, *15* (7), 2958−2973.

(19) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated Computational Discovery of High-Performance Materials for Organic Photovoltaics by Means of Cheminformatics. *Energy Environ. Sci.* **2011**, *4* (12), 4849−4861.

(20) Sahu, H.; Rao, W.; Troisi, A.; Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Adv. Energy Mater.* **2018**, *8* (24), 1801032.

(21) Lee, M.-H. Performance and Matching Band Structure Analysis of Tandem Organic Solar Cells Using Machine Learning Approaches. *Energy Technol.* **2020**, *8* (3), 1900974.

(22) Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine Learning Property Prediction for Organic Photovoltaic Devices. *npj Comput. Mater.* **2020**, *6* (1), 166.

(23) Lee, M. H. Robust Random Forest Based Non-Fullerene Organic Solar Cells Efficiency Prediction. *Org. Electron.* **2020**, *76*, 105465.

(24) Zhao, Z.; del Cueto, M.; Geng, Y.; Troisi, A. Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells. *Chem. Mater.* **2020**, *32* (18), 7777−7787.

(25) Kranthiraja, K.; Saeki, A. Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells. *Adv. Funct. Mater.* **2021**, *31* (23), 2011168.

(26) Zhang, Q.; Zheng, Y. J.; Sun, W.; Ou, Z.; Odunmbaku, O.; Li, M.; Chen, S.; Zhou, Y.; Li, J.; Qin, B.; Sun, K. High-Efficiency Non-Fullerene Acceptors Developed by Machine Learning and Quantum Chemistry. *Adv. Sci.* **2022**, *9* (6), 2104742.

(27) Mahmood, A.; Irfan, A.; Wang, J.-L. Machine Learning and Molecular Dynamics Simulation-Assisted Evolutionary Design and Discovery Pipeline to Screen Efficient Small Molecule Acceptors for PTB7-Th-Based Organic Solar Cells with over 15% Efficiency. *J. Mater. Chem. A* **2022**, *10* (8), 4170−4180.

(28) Hao, T.; Leng, S.; Yang, Y.; Zhong, W.; Zhang, M.; Zhu, L.; Song, J.; Xu, J.; Zhou, G.; Zou, Y.; Zhang, Y.; Liu, F. Capture the High-Efficiency Non-Fullerene Ternary Organic Solar Cells Formula by Machine-Learning-Assisted Energy-Level Alignment Optimization. *Patterns* **2021**, *2* (9), 100333.

(29) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun, K. Machine Learning−Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials. *Sci. Adv.* **2019**, *5* (11), eaay4275.

(30) Peng, S. P.; Zhao, Y. Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors. *J. Chem. Inf. Model.* **2019**, *59* (12), 4993−5001.

(31) Chen, F.-C. Virtual Screening of Conjugated Polymers for Organic Photovoltaic Devices Using Support Vector Machines and Ensemble Learning. *Int. J. Polym. Sci.* **2019**, *2019* (Ml), 1−7.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

(32) Padula, D.; Troisi, A. Concurrent Optimization of Organic Donor−Acceptor Pairs through Machine Learning. *Adv. Energy Mater.* **2019**, *9* (40), 1902463.
(33) Padula, D.; Simpson, J. D.; Troisi, A. Combining Electronic and Structural Features in Machine Learning Models to Predict Organic Solar Cells Properties. *Mater. Horizons* **2019**, *6* (2), 343−349.
(34) Lee, M. Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. *Adv. Energy Mater.* **2019**, *9* (26), 1900891.
(35) Sahu, H.; Ma, H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *J. Phys. Chem. Lett.* **2019**, *10* (22), 7277−7284.
(36) Wu, Y.; Guo, J.; Sun, R.; Min, J. Machine Learning for Accelerating the Discovery of High-Performance Donor/Acceptor Pairs in Non-Fullerene Organic Solar Cells. *npj Comput. Mater.* **2020**, *6* (1), 120.
(37) Lee, M.-H. A Machine Learning−Based Design Rule for Improved Open-Circuit Voltage in Ternary Organic Solar Cells. *Adv. Intell. Syst.* **2020**, *2* (1), 1900108.
(38) Malhotra, P.; Biswas, S.; Chen, F.-C.; Sharma, G. D. Prediction of Non-Radiative Voltage Losses in Organic Solar Cells Using Machine Learning. *Sol. Energy* **2021**, *228*, 175−186.
(39) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **2017**, *1* (4), 857−870.
(40) Liu, X.; Shao, Y.; Lu, T.; Chang, D.; Li, M.; Lu, W. Accelerating the Discovery of High-Performance Donor/Acceptor Pairs in Photovoltaic Materials via Machine Learning and Density Functional Theory. *Mater. Des.* **2022**, *216*, 110561.
(41) Wadsworth, A.; Moser, M.; Marks, A.; Little, M. S.; Gasparini, N.; Brabec, C. J.; Baran, D.; McCulloch, I. Critical Review of the Molecular Design Progress in Non-Fullerene Electron Acceptors towards Commercially Viable Organic Solar Cells. *Chem. Soc. Rev.* **2019**, *48* (6), 1596−1625.
(42) Rodríguez-Martínez, X.; Pascual-San-José, E.; Campoy-Quiles, M. Accelerating Organic Solar Cell Material's Discovery: High-Throughput Screening Andbig Data. *Energy Environ. Sci.* **2021**, *14* (6), 3301−3322.
(43) Sun, W.; Li, M.; Li, Y.; Wu, Z.; Sun, Y.; Lu, S.; Xiao, Z.; Zhao, B.; Sun, K. The Use of Deep Learning to Fast Evaluate Organic Photovoltaic Materials. *Adv. Theory Simulations* **2019**, *2* (1), 1800116.
(44) Zhao, Z.-W.; del Cueto, M.; Troisi, A. Limitations of Machine Learning Models When Predicting Compounds with Completely New Chemistries: Possible Improvements Applied to the Discovery of New Non-Fullerene Acceptors. *Digit. Discovery* **2022**, *1* (3), 266−276.
(45) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(46) Balachandran, P. V.; Xue, D.; Theiler, J.; Hogden, J.; Lookman, T. Adaptive Strategies for Materials Design Using Uncertainties. *Sci. Rep.* **2016**, *6*, 1−9.
(47) Farache, D. E.; Verduzco, J. C.; McClure, Z. D.; Desai, S.; Strachan, A. Active Learning and Molecular Dynamics Simulations to Find High Melting Temperature Alloys. *Comput. Mater. Sci.* **2022**, *209* (March), 111386.
(48) Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; Meredig, B. High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates. *Integr. Mater. Manuf. Innov.* **2017**, *6* (3), 207−217.
(49) Verduzco, J. C.; Marinero, E. E.; Strachan, A. An Active Learning Approach for the Design of Doped LLZO Ceramic Garnets for Battery Applications. *Integr. Mater. Manuf. Innov.* **2021**, *10* (2), 299−310.
(50) Doan, H. A.; Agarwal, G.; Qian, H.; Counihan, M. J.; Rodríguez-López, J.; Moore, J. S.; Assary, R. S. Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chem. Mater.* **2020**, *32* (15), 6338−6346.
(51) Abroshan, H.; Kwak, H. S.; An, Y.; Brown, C.; Chandrasekaran, A.; Winget, P.; Halls, M. D. Active Learning Accelerates Design and Optimization of Hole-Transporting Materials for Organic Electronics. *Front. Chem.* **2022**, *9*, 1−7.
(52) Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Adv. Mater.* **2020**, *32* (14), 1907801.
(53) Kunkel, C.; Margraf, J. T.; Chen, K.; Oberhofer, H.; Reuter, K. Active Discovery of Organic Semiconductors. *Nat. Commun.* **2021**, *12* (1), 2422.
(54) Sun, W.; Zheng, Y.; Zhang, Q.; Yang, K.; Chen, H.; Cho, Y.; Fu, J.; Odunmbaku, O.; Shah, A. A.; Xiao, Z.; Lu, S.; Chen, S.; Li, M.; Qin, B.; Yang, C.; Frauenheim, T.; Sun, K. Artificial Intelligence Designer for Highly-Efficient Organic Photovoltaic Materials. *J. Phys. Chem. Lett.* **2021**, *12* (36), 8847−8854.
(55) Sahu, H.; Yang, F.; Ye, X.; Ma, J.; Fang, W.; Ma, H. Designing Promising Molecules for Organic Solar Cells via Machine Learning Assisted Virtual Screening. *J. Mater. Chem. A* **2019**, *7* (29), 17480−17488.
(56) Zhang, J.; Liu, W.; Zhang, M.; Liu, Y.; Zhou, G.; Xu, S.; Zhang, F.; Zhu, H.; Liu, F.; Zhu, X. Revealing the Critical Role of the HOMO Alignment on Maximizing Current Extraction and Suppressing Energy Loss in Organic Solar Cells. *iScience* **2019**, *19*, 883−893.
(57) Sharma, R.; Jain, N.; Lee, H.; Kabra, D.; Yoo, S. Comprehensive and Comparative Analysis of Photoinduced Charge Generation, Recombination Kinetics, and Energy Losses in Fullerene and Nonfullerene Acceptor-Based Organic Solar Cells. *ACS Appl. Mater. Interfaces* **2020**, *12* (40), 45083−45091.
(58) Menke, S. M.; Ran, N. A.; Bazan, G. C.; Friend, R. H. Understanding Energy Loss in Organic Solar Cells: Toward a New Efficiency Regime. *Joule* **2018**, *2* (1), 25−35.
(59) Teichler, A.; Eckardt, R.; Hoeppener, S.; Friebe, C.; Perelaer, J.; Senes, A.; Morana, M.; Brabec, C. J.; Schubert, U. S. Combinatorial Screening of Polymer: Fullerene Blends for Organic Solar Cells by Inkjet Printing. *Adv. Energy Mater.* **2011**, *1* (1), 105−114.
(60) Wang, J. L.; Liu, K. K.; Hong, L.; Ge, G. Y.; Zhang, C.; Hou, J. Selenopheno[3,2-b]Thiophene-Based Narrow-Bandgap Nonfullerene Acceptor Enabling 13.3% Efficiency for Organic Solar Cells with Thickness-Insensitive Feature. *ACS Energy Lett.* **2018**, *3* (12), 2967−2976.
(61) Pascual-San-José, E.; Rodríguez-Martínez, X.; Adel-Abdelaleim, R.; Stella, M.; Martínez-Ferrero, E.; Campoy-Quiles, M. Blade Coated P3HT:Non-Fullerene Acceptor Solar Cells: A High-Throughput Parameter Study with a Focus on up-Scalability. *J. Mater. Chem. A* **2019**, *7* (35), 20369−20382.
(62) Dong, J.; Cao, D.-S.; Miao, H.-Y.; Liu, S.; Deng, B.-C.; Yun, Y.-H.; Wang, N.-N.; Lu, A.-P.; Zeng, W.-B.; Chen, A. F. ChemDes: An Integrated Web-Based Platform for Molecular Descriptor and Fingerprint Computation. *J. Cheminform.* **2015**, *7* (1), 60.
(63) Miyake, Y.; Saeki, A. Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J. Phys. Chem. Lett.* **2021**, *12* (51), 12391−12401.
(64) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(65) Getting Started with the RDKit in Python -- The RDKit 2021.09.1 documentation. https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors (accessed 2022−03−02).
(66) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **1998**, *13* (4), 455−492.
(67) Vazquez, E.; Bect, J. Convergence Properties of the Expected Improvement Algorithm with Fixed Mean and Covariance Functions. *J. Stat. Plan. Inference* **2010**, *140* (11), 3088−3095.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

(68) Brochu, E.; Cora, V. M.; de Freitas, N.A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv (Machine Learning)*, Dec. 12, 2010, arXiv:1012.2599, version 1, https://arxiv.org/pdf/1012.2599.
(69) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(70) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10* (1), 4.

https://doi.org/10.1021/acsami.2c18540
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX