Article

Accelerating Materials Discovery for Polymer Solar Cells: Data-Driven Insights Enabled by Natural Language Processing

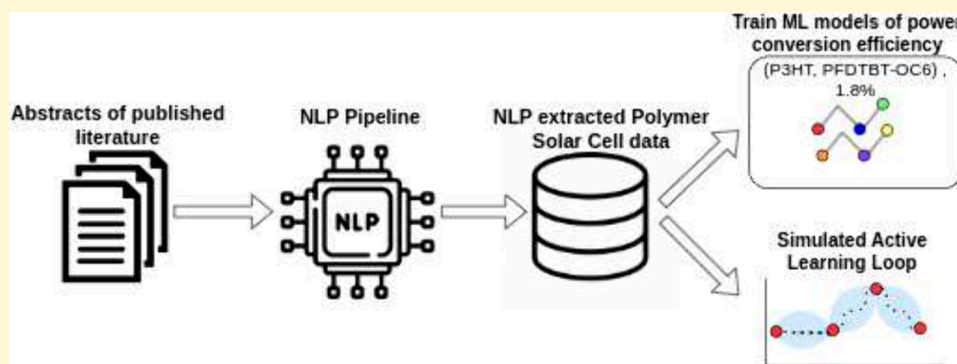Pranav Shetty, Aishat Adeboye, Sonakshi Gupta, Chao Zhang, and Rampi Ramprasad*

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present a simulation of various active learning strategies for the discovery of polymer solar cell donor/acceptor pairs using data extracted from the literature spanning ~20 years by a natural language processing pipeline. While data-driven methods have been well established to discover novel materials faster than Edisonian trial-and-error approaches, their benefits have not been quantified for material discovery problems that can take decades. Our approach demonstrates a potential reduction in discovery time by approximately 75%, equivalent to a 15 year acceleration in material innovation. Our pipeline enables us to extract data from greater than 3300 papers which is ~5 times larger and therefore more diverse than similar data sets reported by others. We also trained machine learning models to predict the power conversion efficiency and used our model to identify promising donor—acceptor combinations that are as yet unreported. We thus demonstrate a pipeline that goes from published literature to extracted material property data which in turn is used to obtain data-driven insights. Our insights include active learning strategies that can be used to train strong predictive models of material properties or be robust to the initial material system used. This work provides a valuable framework for data-driven research in materials science.

INTRODUCTION

Machine learning (ML) methods have become ubiquitous in materials science. Indeed, data-driven methods have enabled the discovery of new materials for applications such as high-breakdown strength dielectric polymers,[1−3] heussler alloys,[4,5] organic photovoltaics with high power conversion efficiency,[6,7] and gas separation membranes with high selective permeability.[8] Active learning is a technique that is used to drive this improvement. It leverages trained models of material properties to identify promising candidate materials and then augment the training set once the candidate material is "measured", to improve the predictive performance of the ML models. Active learning methods have been used to discover high hole mobility thin films,[9] alloys for gas turbine engine blades,[10] and high glass transition temperature polymers.[11] Recent work has benchmarked the performance of various active learning strategies on problems such as electro-catalysis,[12] magnetic properties,[13] and band gap.[14] Several studies have also benchmarked the performance of human scientists against bayesian optimization algorithms for the

optimization of materials systems such as polyoxometallate clusters[15,16] and palladium-catalyzed arylations.[17] However, these studies were laboratory-scale experiments that were run under a controlled setting over a short time scale relative to the many decades it can take for materials discovery to play out for important applications. Polymer solar cells offer a promising playground for us to quantify the time savings offered by the use of machine learning methods in materials science using historical data.

Traditionally solar cells are inorganic and are made of crystalline silicon or amorphous silicon. Light incident on the solar cell's surface produces electron—hole pairs, generating a

A

potential that can drive an external load. Due to the use of an inorganic material like silicon, the cost of manufacturing such cells is increased due to the high temperatures required to process silicon. Polymer solar cells, on the other hand, are made of organic materials and can thus be processed at much lower temperatures. The most common configuration of a polymer solar cell is as a bulk heterojunction blend of two materials known as the donor and acceptor. Finding optimal donor/acceptor combinations that lead to high power conversion efficiency (PCE) is an active area of research. Historically, the acceptor was based on fullerenes, typically [6,6] phenyl-C61-butyric acid methyl ester (PCBM). In recent years, though nonfullerene acceptors have gained traction and have led to solar cell configurations with much higher efficiency.[18]

The chemical space for donors and acceptors is vast, making machine learning methods a promising avenue for narrowing down the chemical space before testing candidate materials. Past work has developed machine learning models for polymer solar cells to identify promising donors for fullerene systems[19] as well as promising nonfullerene acceptors.[20] Ref 7 uses trained ML models of PCE to find promising candidate materials and then tests cells fabricated using ML predictions to discover new viable candidates. The number of experimental papers published in this space is quite large however and covers a wide chemical space. In this work, we found over ~3300 relevant papers using our pipeline. The papers that typically applied ML methods to polymer solar cells relied on data that was manually collected and covered about 500 papers on average. A summary of publicly available polymer solar cell data sets is provided in Table 1. Thus, due to the manual

Table 1. Comparing Public Data Sets of Polymer Solar Cell Device Characteristics[a]

Type of data	Number of unique data points	Number of papers	Reference
Fullerenes only	1200	500	19
Nonfullerenes only	1001	—	21
Nonfullerenes only	1318	558	6, 20
Fullerenes only	350	—	22
Fullerenes and nonfullerenes	PSC_{NLP}: 2585; $PSC_{Curated}$: 867	PSC_{NLP}: 3307; $PSC_{Curated}$: 861	This work

[a]PSC in this table refers to PolymerSolarCells.

nature of data collection, nearly five-sixth of relevant papers typically cannot be parsed for data extraction due to time constraints. The use of natural language processing (NLP) offers a potential solution to the problem of collecting material property data at scale.

NLP is a way to get computers to understand human text. NLP has been used in materials science to obtain insights from inorganic literature as well as polymer literature by extracting structured material property data or synthesis data from a large number of papers.[23−32] We used NLP methods to extract device characteristics of polymer solar cells from the abstracts of materials science literature. A data extraction pipeline was developed as part of a previous work[33−35] that recognized the categories of words in text using machine learning models and used heuristic rules to link them into a material property record. Using this pipeline, we obtained a data set of donors and acceptors and their corresponding device characteristics such as PCE which we call PolymerSolarCells$_{NLP}$. We curated a

subset of PolymerSolarCells$_{NLP}$ to record the structure of the donor and acceptor as a SMILES string[36] which we call PolymerSolarCells$_{Curated}$.

Using PolymerSolarCells$_{Curated}$, we trained a machine learning model that takes the donor and acceptor as input and predicts the PCE (overall pipeline shown in Figure 1). We used this trained model to predict the PCE for all donor/acceptor combinations not found in PolymerSolarCells$_{Curated}$ to identify promising candidates with high PCE. This training methodology was also used while simulating an active learning loop. The fundamental premise of using machine learning in materials science is that we can discover new material systems faster than would otherwise be possible through trial-and-error. We provide the first quantitative evidence for this that goes beyond laboratory scale active learning experiments by comparing the sequence of donor/acceptor systems obtained using active learning to the actual evolution of PCE for polymer solar cells over 20 years. We thus make a stronger policy case for wider systemic adoption of machine learning for the discovery of new materials and provide some best practices for the deployment of active learning for materials discovery. All code and data needed to reproduce the results in this paper can be found at https://github.com/pranav-s/PolymerSolarCellsML.

■ METHODS

Data Extraction Pipeline from Literature. We created a corpus of 2.4 million materials science journal articles by web scraping and by downloading from publisher-specific APIs. Elsevier, Wiley, Royal Society of Chemistry, American Chemical Society, Springer Nature, Taylor & Francis, and the American Institute of Physics are the publishers included in our corpus.[33] These papers span the years 2000 to 2022. A subset of 750 abstracts was annotated using an ontology specific to the materials science domain consisting of the entity types POLYMER, POLYMER_CLASS, PROPERTY_VALUE, PROPERTY_NAME, MONOMER, ORGANIC_MATERIAL, INORGANIC_MATERIAL, and MATERIAL_AMOUNT. A named entity recognition (NER) model was trained using the annotated data set. The trained NER model along with heuristic rules for entity linking was used to extract material property data from the abstracts of all papers in our corpus that were polymer-relevant. This extracted data set represents all material property data reported in the abstracts of our corpus of papers. A subset of this material property data corresponds to polymer solar cells. The selection and curation of this subset are described next. Refer to ref 35 for further details.

Creating PolymerSolarCells$_{NLP}$. Polymer solar cells was the domain with the largest number of papers in our corpus. (Note that our usage of polymer solar cells refers to cells in which at least the donor is a polymer,[37−39] in contrast to all-polymer cells[40,41] in which the donor and acceptor are polymers.) It is conventional for scientists to record information such as the donor/acceptor system developed in the paper as well as its key device characteristics in the abstract of the paper. Moreover, this field has for the most part developed over the ~20 year time span between 2000 and 2022 which overlaps well with the time frame of our corpus of papers. Most of the research in this field has also focused on developing a donor/acceptor that improves one key property, namely the power conversion efficiency of the resulting cell. This makes it an ideal playground for our active learning experiments.

We picked the subset of abstracts (from which we extracted all material property data) that contained the keyword "solar cell" and contained reported values for PSC device characteristics. We also excluded keywords such as "perovskite", "dye-sensitized", "tandem solar cell", "quantum dot", "hybrid solar cell", "silicon solar cell", and "ternary solar cell". In practice, it is common for authors working in polymer solar cells to report in the abstract the device characteristics associated with the most important donor−acceptor combinations
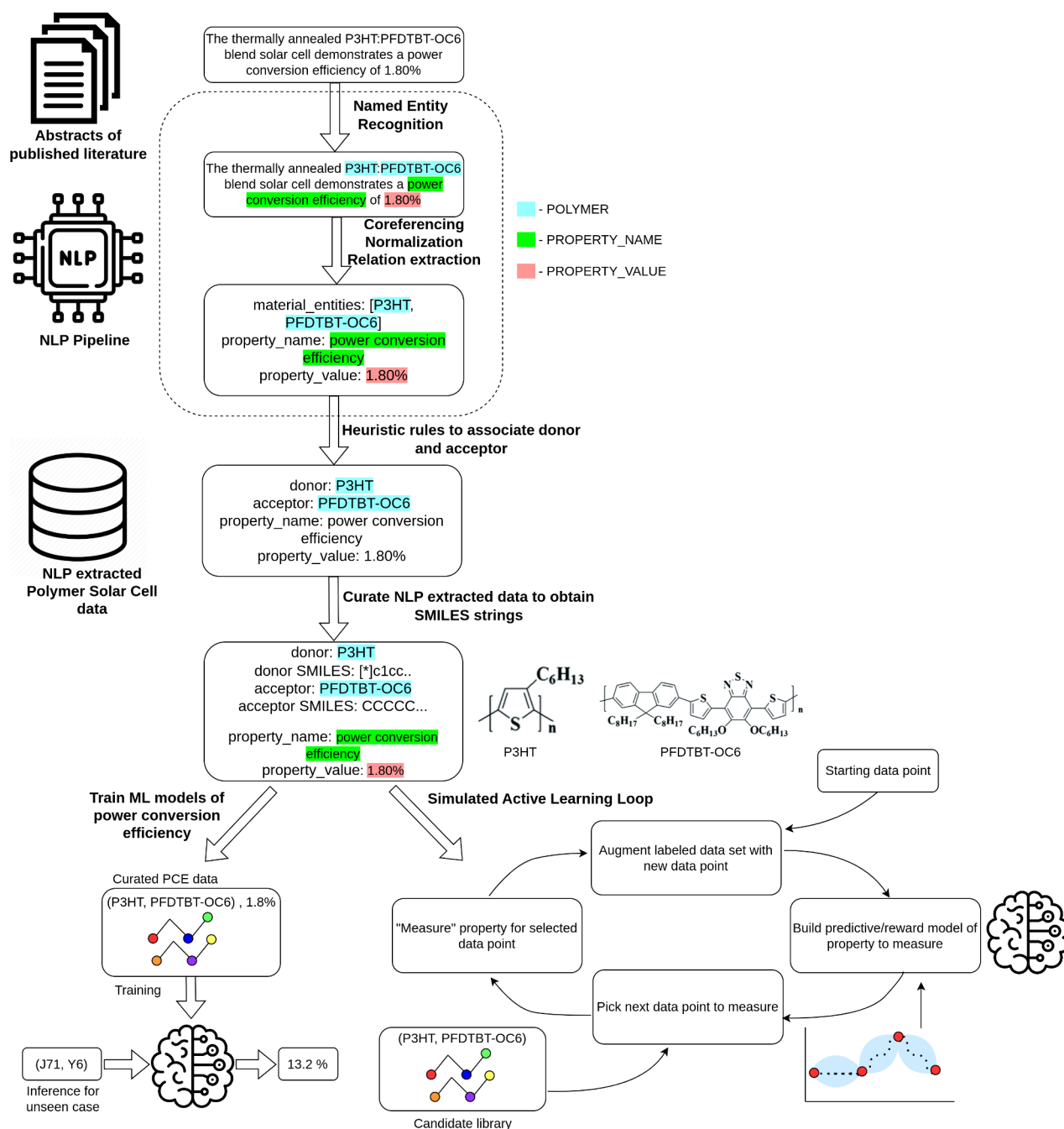
Figure 1. Pipeline used for extracting polymer solar cell PCE data from published literature which is then used in two ways: (1) To train ML models of PCE and predict high-performing donor/acceptor pairs not reported in the literature and (2) to simulate an active learning loop through which donor/acceptor pairs are "discovered" sequentially. (J71, Y6) referenced in the figure is a donor/acceptor pair.

tested in the paper, as they are the most significant results reported. Since the ontology used in ref 35 was not specific to polymer solar cells, it was necessary to distinguish donor and acceptor labels for material entities reported in the abstract. We labeled material entities as donors or acceptors based on the following hierarchy of rules. (1) Donors and acceptors are often separated by a "/" or a ":" such as "P3HT/PCBM" or "P3HT:PCBM", wherein the first entry is the donor and the second entry is the acceptor. Material entities mentioned in this manner were labeled as donors and acceptors, respectively. (2) POLYMER or ORGANIC entities that co-occur with the word donor or acceptor were labeled as the donor and acceptor, respectively. (3) If PCBM or any other fullerene is mentioned in the text then it was labeled as the acceptor. (4) If an ORGANIC entity is present in the list of materials for a material property record then it

was labeled as the acceptor and the POLYMER entity in that material list was labeled as the donor. This case could also correspond to a small molecule donor and a polymeric acceptor. However, small molecule donors are rarely observed in our data set and are hence ignored in this rule. Anything that fell outside the above set of rules was marked as being ambiguous and required manual curation.

It is common for polymers to be referred to by generic identifiers such as P1 or P2 and for the corresponding structure to be provided in the paper. In such cases, we changed the name of the entity to P1_DOI where DOI is the digital object identifier of the paper. This helps distinguish this case from other papers where similar identifiers were used. PolymerSolarCells$_{NLP}$ thus contains NLP extracted donor/ acceptor pairs and their corresponding device characteristics. The

NLP pipeline also extracts coreferents (such as abbreviations) for the donor and the acceptor.

Curating PolymerSolarCells$_{NLP}$ to create PolymerSolarCells$_{Curated}$. To train machine learning models, we curated a subset of PolymerSolarCells$_{NLP}$ to (1) fix any extraction errors made by the NLP pipeline and (2) collect SMILES strings for donors and acceptors which can be used to generate a structural fingerprint of the material entity for training ML models. The SMILES string is a string representation of the structure of a molecule.

With each extracted data point in PolymerSolarCells$_{NLP}$ as the starting point, curators checked the corresponding abstract to see if the NLP extracted data was correct. If any part of the NLP extracted data was incorrect, then annotators fixed the entities extracted incorrectly. This included ensuring that the donor, acceptor, and corresponding device characteristics mentioned in the abstract, i.e., PCE, fill factor, open circuit voltage, and short circuit were recorded correctly whenever present. Any relevant material property data present in the abstract that the NLP pipeline failed to extract was recorded manually by the annotators. There were cases where only the donor or the acceptor was mentioned in the abstract while the other had to be located in the body of the paper. In such cases, the material entity not found in the abstract is provided in single quotes to distinguish it from cases where the material entity is provided in the abstract. The SMILES strings for all donors and acceptors were also recorded. The SMILES strings for the most common donors and acceptors found in our corpus were first created and then prepopulated in the corresponding data point to avoid repeating the effort of recreating those SMILES strings. For every other case, the SMILES string was constructed by looking for the structure among the figures in the paper and using the draw tool found in polymergenome.org to construct the SMILES string. In cases where the structure was not found in the paper, the annotators looked for supporting references that contained the structure. Note that we used the psmiles convention as defined in ref 42 for polymer SMILES strings. Creating the SMILES strings was the most time-consuming aspect of data curation.

In addition to the basic curation workflow described above, we employed several filtering criteria. We were only interested in properties reported for bulk heterojunction polymer solar cells. Thus, the donor in each data point is a polymer and papers dealing with ternary systems, i.e., with multiple donors or multiple acceptors, were excluded. This was done so that the learning problem could be formulated using a single donor and acceptor. Small molecule donors were excluded as very few papers in PolymerSolarCells$_{NLP}$ report such data given that our NLP pipeline was engineered to extract polymer property data. Out of 150 papers randomly sampled from PolymerSolarCells$_{NLP}$, there was only one paper corresponding to a small molecule donor. We also excluded papers reporting properties computed through simulations or materials that were discovered using data-driven methods such as ref 7.

The device characteristics reported in PolymerSolarCells$_{Curated}$ are power conversion efficiency, fill factor, open circuit voltage, and short circuit current. In the analysis that follows, we only make use of the power conversion efficiency. Note that all property values curated were mentioned in the abstract. In the case that multiple property values are found in the abstract corresponding to different fabrication conditions, the annotators were instructed to record only those corresponding to the highest power conversion efficiency. This corresponds to the optimal fabrication condition.

Machine Learning Prediction of Power Conversion Efficiency. We used PolymerSolarCells$_{Curated}$ to build machine learning models of the power conversion efficiency (PCE) of a polymer solar cell. We modeled the PCE as being dependent on only the donor and acceptor. In practice, the PCE depends on factors like the weight fractions of the donor and acceptor and other device fabrication parameters.[43] We however assumed that the property values reported in the abstract are for systems in which other factors have been optimized. This is because scientists typically report the highest PCE value they measure as one of the key contributions of their paper which in turn corresponds to an optimized cell. We empirically

validated this assumption by sampling 20 papers in PolymerSolarCells$_{Curated}$ and verified that this assumption held true in all 20 papers. Our problem formulation is more general than several others reported in the literature which only use the structure of the donor and assume that the acceptor is a fullerene[19] or do not consider the effect of the acceptor.[7] Due to the recent rise of nonfullerenes,[18] it is necessary to include acceptors in the problem formulation as done in ref 44. For donor−acceptor pairs for which multiple power conversion efficiencies are available, we took the median of all available values. The values may differ due to differences in the electron or hole transport layer, additives used during the synthesis of the active layer, or the morphology of the active layer. Taking the median "averages" out these factors.

We used hand-crafted fingerprints described in earlier work[45] to fingerprint the donor and acceptor. While other fingerprints have been developed for featurizing polymers using ideas such as graph neural networks[46] and BERT-based fingerprinting,[42] we used handcrafted fingerprints for this work as they can be used to fingerprint polymers as well as organic molecules. Moreover, each fingerprint component corresponds to a predefined chemical feature which enables us to "peer" into the model and determine why the model made certain predictions. In cases where the donor or acceptor is a copolymer, we take the average fingerprint of all components of the copolymer, in keeping with past work.[47] Further details on fingerprinting methodology can be found in the Supporting Information, Section 2. Min-max scaling is performed on all fingerprint components to ensure that they are each between 0 and 1. The vector fingerprints for the donor and acceptor were then concatenated and input to a Gaussian Process Regression (GPR) model to predict the power conversion efficiency. The fingerprint dimension is 745 while the number of data points is 835. (This is lower than the total number of unique donor/acceptor systems (867) contained in PolymerSolarCells$_{Curated}$ as 32 donor/acceptor systems in PolymerSolarCells$_{Curated}$ have device characteristics other than PCE such as fill factor or open circuit voltage reported in the corresponding abstracts.) Neural networks tend to overfit when the number of data points is close to the dimension of the feature space while GPR models generalize much better in this regime, making it a natural choice for our problem.[48] We used the Python package scikit-learn[49] to train GPR models. We used a Matern kernel[50] with ν = 1.5. We chose this kernel as it is well tested in a variety of materials informatics problems.[51−53] PolymerSolarCells$_{Curated}$ is split into 85% for training and 15% for testing. We used the root mean squared error (RMSE) and the coefficient of determination (r) to assess model performance.

Data Selection Methods for Simulating Active Learning of Polymer Solar Cells. Using PolymerSolarCells$_{Curated}$ as the set of candidate materials, we picked a sequence of donor/acceptor pairs using several data selection methods. This process was continued until the data point with the highest PCE in the data set was discovered. This in effect simulates an active learning generated path of how the field of polymer solar cells may have developed and can be used to compare against how the field developed. Let y_i^{pred} be the predicted value of the GPR model for the ith test point and σ_i be the corresponding uncertainty of the GPR model evaluated at that point where $i \in \{1, ..., N\}$ and N is the number of test points at a given step in the active learning loop. Suppose the current step in the active learning cycle is T. Let Y_i represent the random variable corresponding to the measurement of the ith test point. If GPR is used to model Y_i, then it is known that Y_i is Gaussian distributed with mean y_i^{pred} and standard deviation σ_i. We employed several different methods to pick the next data point to test: (1) **Gaussian Process-Upper Confidence Bound (GP-UCB)**: This uses the predicted value along with its uncertainty to pick points, i.e., $y_i^{\text{pred}} + \beta\sigma_i$. The parameter β controls the trade-off of exploration versus exploitation. A higher value of β indicates a belief that higher value points are likely to be found in regions of uncertainty while a lower value of beta indicates a belief that the model predictions should be exploited as is. We used a value of $\beta = 1$ in this work as it represents a balanced trade-off of exploration and exploitation.[54] (2) **Gaussian Process-**

Probability Improvement (GP-PI): Intuitively, this method looks for points that have the highest probability of being greater than some threshold θ. $\mathbb{P}(Y_i > \theta) = 1 - \Phi\left(\frac{\theta - y_i^{\text{pred}}}{\sigma_i}\right) = \Phi\left(\frac{y_i^{\text{pred}} - \theta}{\sigma_i}\right)$. Here $\Phi(.)$ is the cumulative probability distribution function for a standard Gaussian. In practice, the threshold is selected to be the value of the highest y sampled until that point, multiplied by some improvement fraction $1 + \nu$, i.e., $\theta = \max_{t \in \{1, \cdots, T\}} y_t (1 + \nu)$. We pick a value of $\nu = 0.01$. This provides a practical compromise between being too aggressive or too conservative in seeking improvements across all active learning cycles. This is enough to avoid negligible improvements between cycles but is not so large as to miss potential subtle improvements in areas close to the current best observations thus leading to consistent and gradual improvement.[55] After computing the probability of improvement for all test points, we pick the point with the greatest value. (3) **Gaussian process-expected improvement (GP-EI)**: In contrast to the previous method where we compute the probability of improvement over some threshold, here we look at the expected value of the difference between the value and the threshold. Thus, we compute the quantity $\mathbb{E}[Y_i - \theta] = \int_{-\infty}^{\infty} (y_i - \theta) p(y_i) \, dy_i$.

This evaluates to $E[Y_i - \theta] = (y_i^{\text{pred}} - \theta)\Phi\left(\frac{y_i^{\text{pred}}}{\sigma_i}\right) + \sigma_i \phi\left(\frac{y_i^{\text{pred}}}{\sigma_i}\right)$, where $\phi(.)$ is the probability distribution function of the standard Gaussian. The threshold is computed the same way as GP-PI. After computing this quantity for all test points, we pick the point with the largest expected improvement. (4) **Greedy acquisition**: This method simply looks at the highest predicted values y_i^{pred}. This is equivalent to using $\beta = 0$ in GP-UCB. (5) **Gaussian Process-Thompson Sampling (GP-TS)**: For each material system, we sample the value of the power conversion efficiency from a Gaussian distribution with mean y_i^{pred} and standard deviation σ_i. We pick the material system with the highest sampled property value. (6) **Linear contextual bandits**: In a multiarm bandit setting, an agent interacts with an environment over several rounds. In contrast to the methods described above, we do not train an ML model at each round. During each round, the agent selects an action from a set of available actions (also known as "arms" in the multiarm bandit setting) where each action is represented by a d-dimensional feature vector (the "context"). In this case, the set of actions is the available set of donor/acceptor systems. After taking an action, the agent receives a reward, which in this case is the value of the PCE "measured" for a donor/acceptor system. The goal is to learn a policy (a mapping from context to action) that maximizes the expected cumulative reward over time. In linear contextual bandits, the reward function is assumed to be a linear function of the action. The agent's goal is to estimate the parameters of the linear model based on observed data and use it to make action selections that maximize expected rewards. We implemented Thompson sampling for linear contextual bandits.[56] While training GPR models, the focus is on optimizing prediction accuracy, while in linear contextual bandits, the focus is on maximizing cumulative rewards. Instead of having a fixed data set used for training, contextual bandits involve online learning where the agent interacts with the environment and adapts its policy over time. Further mathematical details are provided in Supporting Information, Section 1. (7) **Random**: A donor/acceptor pair is sampled uniformly from the candidate set at each step of the active learning cycle.

In each of the above methods, we could pick a batch b of any size, but in practice, we use $b = 1$ since the cost of fabricating a cell with a given material system and measuring the PCE is much larger than the computational cost of the active learning loop.

■ RESULTS AND DISCUSSION

Analysis of Polymer Solar Cell Data. There are 3307 documents in PolymerSolarCells$_{\text{NLP}}$ out of which 861 (~26%) have been curated to create PolymerSolarCells$_{\text{Curated}}$ (detailed breakdown in Table 2). The number of unique data points in PolymerSolarCells$_{\text{NLP}}$ is estimated based on the normalized

Table 2. Detailed Composition of the NLP Extracted Polymer Solar Cell Data and the Curated Subset

Type of data	PolymerSolarCells$_{\text{NLP}}$	PolymerSolarCells$_{\text{Curated}}$
Number of papers	3307	861
Unique donor-fullerene pairs	1160	607
Total donor-fullerene data points	2107	903
Unique donor-nonfullerene pairs	1425	261
Total donor-nonfullerene data points	1934	284
Number of unique donors	1910	628
Number of unique acceptors	649	190
Maximum reported power conversion efficiency	–	16.4%

names of the donors and acceptors and is an approximation. The donors are always polymers and hence can be normalized using the normalization workflow described in earlier work.[34] The acceptors on the other hand can be polymers or organic molecules. For polymer acceptors, we used our polymer normalization workflow. For fullerene acceptors, given that only about four such acceptors are commonly used (PC61BM, PC71BM, C60, ICBA), we manually constructed a list of named entity variants for each of these. For other nonfullerene acceptors, we constructed an on-the-fly normalization data set using the coreferents for all nonfullerene acceptors mentioned among the curated papers. The number of total data points is greater than the number of unique data points as there are several material systems reported in multiple papers, usually differing in cell fabrication conditions. For PolymerSolarCells$_{\text{Curated}}$, the number of data points was computed using the canonical SMILES string of the donor and acceptor. The SMILES strings were canonicalized using the psmiles package described in ref 42. Canonicalization is important as the same chemical structure can be represented as a valid SMILES string in multiple ways, and canonicalization is used to create a unique SMILES string for a given chemical structure.

We can see from the heatmap in Figure 2 that only a small fraction (0.70%) of the donor/acceptor space has been explored experimentally. Not all donors and acceptors are labeled as they are far too numerous. The donors on the heat map which are reported against the most acceptors correspond to P3HT (40), PTB7-Th (40), and PBDB-T (37). The material system with the highest reported PCE for a bulk heterojunction polymer solar cell in PolymerSolarCells$_{\text{Curated}}$ is PBDS-T as donor and BTP-ec9 as acceptor with a reported PCE of 16.4%.[57] We built a prediction pipeline that can fill in the blanks in this figure and used that to discern patterns as well as find optimal donor/acceptor combinations that would otherwise require expensive experimental measurements. A machine learning model that takes both the donor and the acceptor structure as input should in principle be able to learn correlations between the two which can be used to predict novel combinations of donors and acceptors.

In Table 3 we compare the ground truth versus the extracted data to measure the fidelity with which the NLP system can extract the relevant polymer solar cell data.

We measured the fidelity of extraction using k-tuple metrics where $2 \leq k \leq 4$. For a 4-tuple comparison, we looked at the tuple (donor, acceptor, property_name, property_value). We compared the ground truth tuples against the NLP extracted
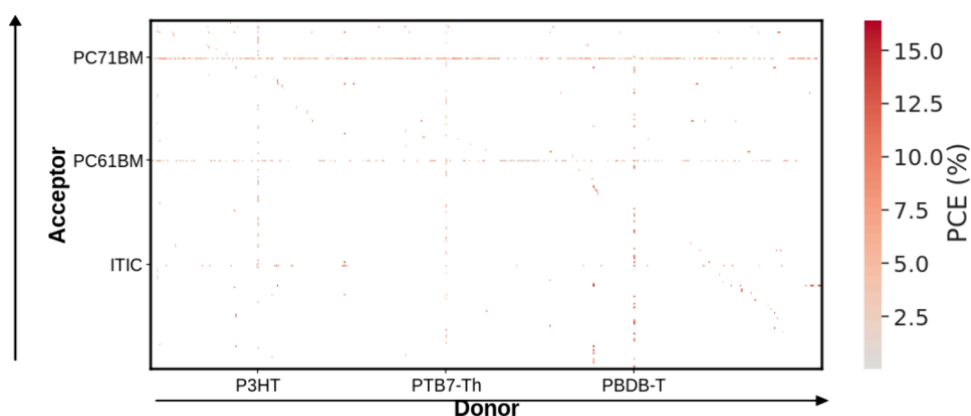
Figure 2. Entire donor/acceptor space at a glance. The donors and acceptors that are most frequently reported are shown along the axes. The top three most commonly reported donors and acceptors are spaced uniformly along each corresponding axes so that they can be clearly distinguished. The remaining donors and acceptors are randomly ordered.

Table 3. Comparing Fidelity of NLP Extracted Data to Ground Truth Data[a]

Metric	Precision	Recall	F1
4-tuple	46.03	40.25	42.95
3-tuple	61.34	55.09	58.05
2-tuple	72.87	66.18	69.37

[a]Each value in the table is a %.

tuples for each abstract to check if there is any NLP extracted tuple for which all 4 entries are identical and that is marked as a true positive (TP). If no such NLP extracted tuple is found, we have a false negative (FN). We similarly compared the NLP extracted tuples against the ground truth tuples for each abstract to check if a match is found and if not, then a false positive (FP) is recorded. As the 4-tuple metric is strict, we computed 3-tuple and 2-tuple metrics as well, wherein the 4-tuple is split into 3 and 6 tuples respectively for each 4-tuple, in each abstract. This allows us to see if at least some subsets of the 4-tuple are extracted correctly. This is consistent with how extraction fidelity has been measured in the literature in cases

involving multiple entities in a tuple.[58] Precision, Recall, and F1 score are calculated as below:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (1)$$

Each of the above metrics is reported as a % value.

Predicting Power Conversion Efficiency. The RMSE of property prediction is 2.07% for the model using donors and acceptors (Figure 3). Figure 3 compares a machine learning model trained using the donor alone and using the donor as well as the acceptor as input to the model. The donor-only data set was constructed by removing the acceptors from each data point in the donor/acceptor data set thus ensuring a fair comparison. Thus, some data points have the same donor but different values of power conversion efficiency which is due to the corresponding acceptors being different. Having additional
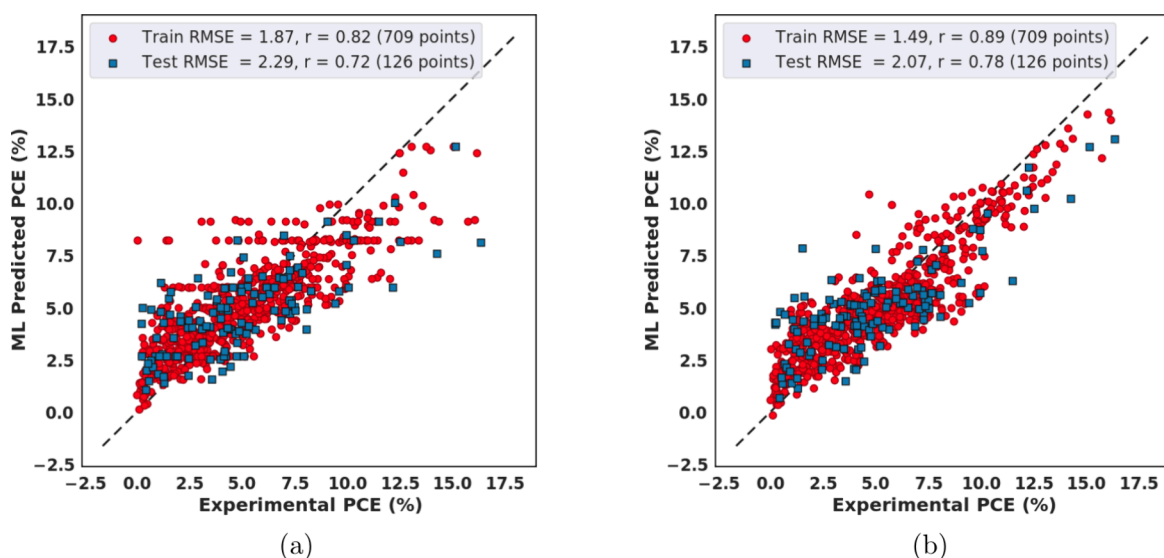

Figure 3. Parity plot for a machine learning model trained to predict power conversion efficiency. (a) Model trained using only donors as input. (b) Model trained using donors and acceptors as input to the model.
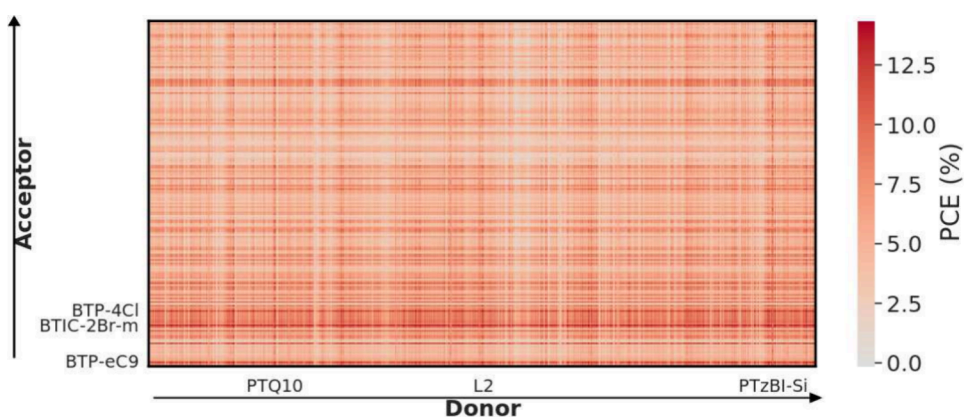
Figure 4. Predicted power conversion efficiency value for the entire donor/acceptor space. The donors and acceptors with the highest average PCE are shown along the corresponding axes. The ordering of donors and acceptors is the same as Figure 2.

information on the acceptor improves the prediction performance of the model by 0.22% as measured by the test RMSE. Note that the train-test split is identical for both panels a and b in Figure 3. Figure 4 shows the predicted power conversion efficiencies for all donor/acceptor pairs in PolymerSolarCells_Curated. We observe from Figure 4 that the use of certain acceptors can improve the power conversion efficiency significantly. The bright red horizontal line at the bottom corresponds to the acceptor BTIC-2Br-m which is a Y6-based acceptor. Similarly, some of the other high-performance acceptors in Figure 4 are also based on Y6[59] such as BTP-4Cl and BTP-ec9. The observations we make from this figure match well with the strategy recently employed in the polymer solar cell community of using acceptors based on Y6 and testing it with various donors.

The highest donor/acceptor pairs predicted from our model are listed in Table 4 along with the corresponding power

Table 4. Donor/Acceptor Pairs with Highest Predicted Power Conversion Efficiency

Donor	Acceptor	Predicted power conversion efficiency	True PCE (if reported in literature)
PBDB-TF	BTIC-2Br-m	14.34%	16.1%[60]
PM6	BTIC-2Br-m	14.33%	–
PFBDB-T[61]	BTIC-2Br-m	14.33%	–
PM7	BTP-ec9	14.31%	–
PM7	BTP-4Cl	14.30%	–
PM7	TPIC-4Cl	14.25%	15.1%[62]
PDBT-F[63]	BTIC-2Br-m	14.20%	–
L2	BTP-ec9	14.17%	–
L2	BTP-4Cl	14.16%	–
PBDT(T)[2F]T[64]	BTIC-2Br-m	14.07%	–

conversion efficiency. All the acceptors are based on Y6, and the corresponding donors are thus recommendations for promising combinations. Due to the sparsity of the underlying data set, however, this extrapolation must be viewed with caution. It is possible that the acceptors with the highest recorded PCE have an undue influence on the predictions and that donor/acceptor interactions cannot be fully captured due to the sparsity of the underlying data. The vast majority of donors are reported with just a single acceptor and, likewise, most acceptors are tested with just a single donor, thus making it difficult for a model to learn the correlations effectively. To

examine this possibility more closely, we looked at the top donors and acceptors from Figure 4 by computing the average PCE over donors for each acceptor and likewise for donors (Table 5). The top pairs reported in Table 4 are not simply

Table 5. Top Donors and Acceptors from Figure 4

Top Donors	Average PCE (%)	Top acceptors	Average PCE (%)
PTzBI-Si[65]	8.87	BTIC-2Br-m[60]	12.42
L2[66]	8.68	BTP-4Cl[67]	11.19
PTQ10[68]	8.46	BTP-ec9	11.07
PM7[69]	8.44	TPIC-4Cl[62]	10.42
F13[70]	7.98	Y6	9.37

formed by matching the top donors with the top acceptors in Table 5 but are nontrivial combinations of donors and acceptors, suggesting that the model may be learning meaningful donor/acceptor corelations. The acceptors used in the top reported pairs match closely with the top acceptors overall, and the average PCE for the top acceptors is higher than that for the top donors suggesting that acceptors do indeed play a more important role than donors in determining the power conversion efficiency in the high PCE regime.

Thus, we have demonstrated the end-to-end use of NLP to facilitate the creation of a data set which is then used for building machine learning models and predicting new donor/ acceptor combinations for polymer solar cells. This also validates our modeling approach for simulated active learning discussed in the next section.

Simulating the "Discovery" of New Donor/Acceptor Combinations. The green line in Figure 5a−c shows how the power conversion efficiency of polymer solar cell systems reported in the literature has changed over time. Each point corresponds to a single donor/acceptor system. In cases where a donor/acceptor system is reported multiple times in the literature, we take the median value and the timestamp for the paper corresponding to the median value (when the number of data points n is even, we use the PCE at index $\frac{n}{2}$ after the PCE values are sorted). Observe that the value of PCE shows an upward trend until 2022 when a peak of 16.4% was obtained which is the highest value in PolymerSolarCells_Curated. Thus, this plot captures a near complete picture of the evolution of the field of polymer solar cells. The discovery of new material systems has proceeded through trial and error which is why the PCE does not increase monotonically but fluctuates in value
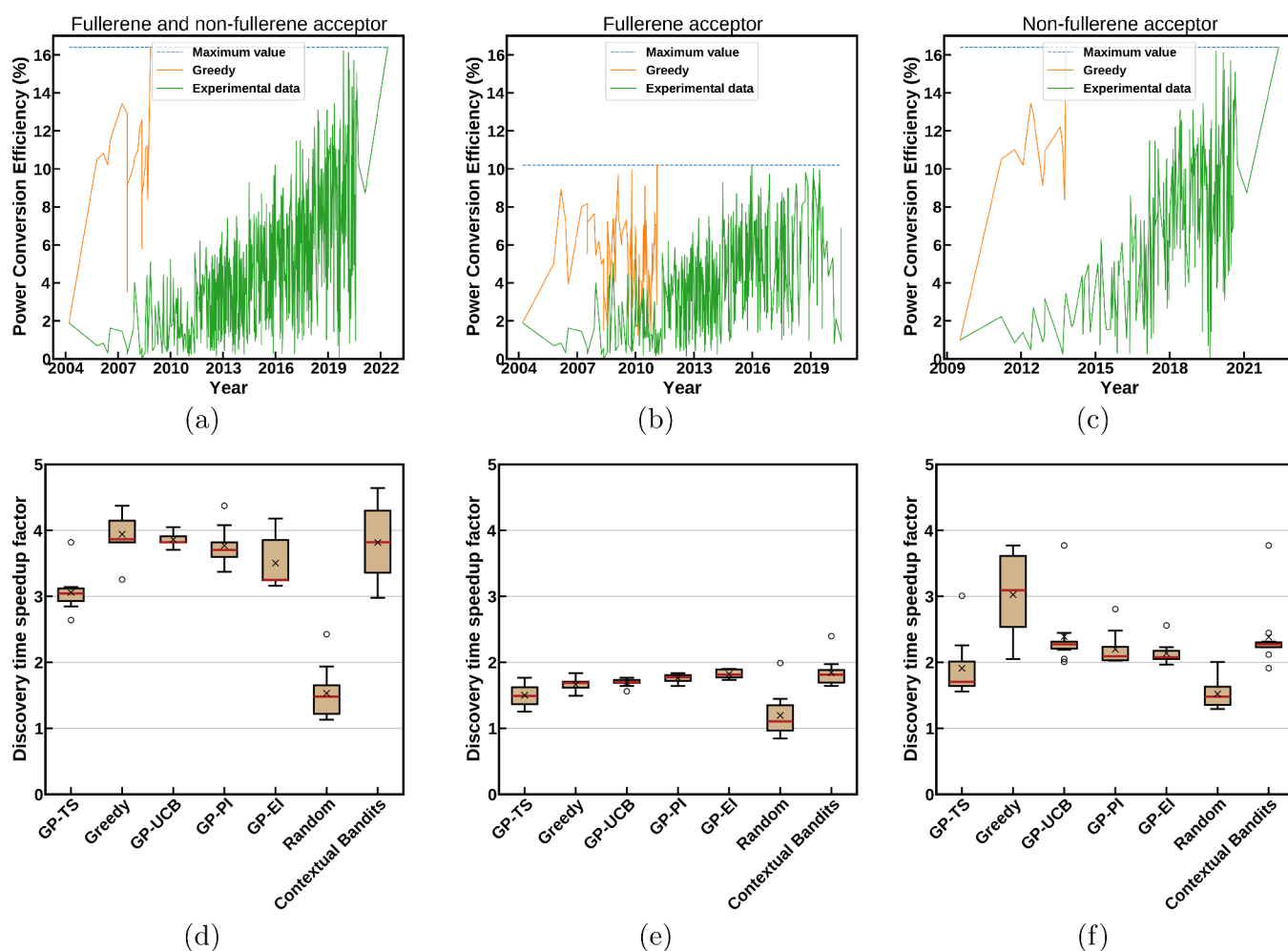
Figure 5. Comparing the simulated path of material systems generated by data selection methods against the evolution of power conversion efficiency in the experimental literature. (a) Both fullerene and nonfullerene acceptors included among candidate material systems, (b) fullerene acceptor only, and (c) nonfullerene acceptors only. (d−f) Range of DTSF obtained using ten different starting material systems for each data selection method tested for the acceptors used in the panel above. In the box and whisker plot in panels (d−f), the red line indicates the median of the data while † represents the mean of the data. O are outliers.

over time. The number of PCE values reported that underperform the then-state-of-the-art is likely underestimated on this plot given that only material systems considered improvements get published. This results in a bias in the available data.[71]

We consider how this field may have evolved had it relied completely on data-driven methods to pick the next donor/acceptor combinations to evaluate rather than relying on trial and error. This allows us to empirically estimate how much faster we could have reached the end-point shown in Figure 5a−c.

Figure 1 shows the iterative process we used for material selection. The initial donor/acceptor system we used was OC_1C_{10}-PPV/$PC_{61}BM$[72] which was the earliest reported material system in PolymerSolarCells$_{Curated}$. At each step, we trained a GPR model using all data already "measured" and used that to make predictions on the remaining donor/acceptor pairs in PolymerSolarCells$_{Curated}$. We picked one donor/acceptor system that was the most "promising" according to a few different methods. We "measured" the true power conversion efficiency by looking up the value from PolymerSolarCells$_{Curated}$. This new data point was then used to augment the training data set to train a new model and repeat

the above cycle. The iterations were terminated when the material system with the highest value in our data set was identified. This is identical to active learning with the key difference being that we can compare active learning generated paths of donor/acceptor pairs against how the field evolved and estimate the speedup. We make three key assumptions in this analysis. (1) The list of donor/acceptor systems used as candidate material systems is the full list of material systems reported in PolymerSolarCells$_{Curated}$. While we do not assume foreknowledge of the actual value of PCE, coming up with viable new material systems was part of the evolution of this field which we are unable to realistically simulate. A fairer analysis would assume a much larger candidate list of donor/acceptor systems which would be created without the bias that they were reported in the polymer solar cell literature. There would however be no way to "measure" the value of PCE for these material systems with the same fidelity as experimentally reported values. (2) The timestamp associated with each new donor/acceptor system that is picked is the same as the timestamp for an entry of the corresponding index in the experimental data. As the experimentally reported results come from many different research groups, this assumption is reasonable if we assume that many different research groups
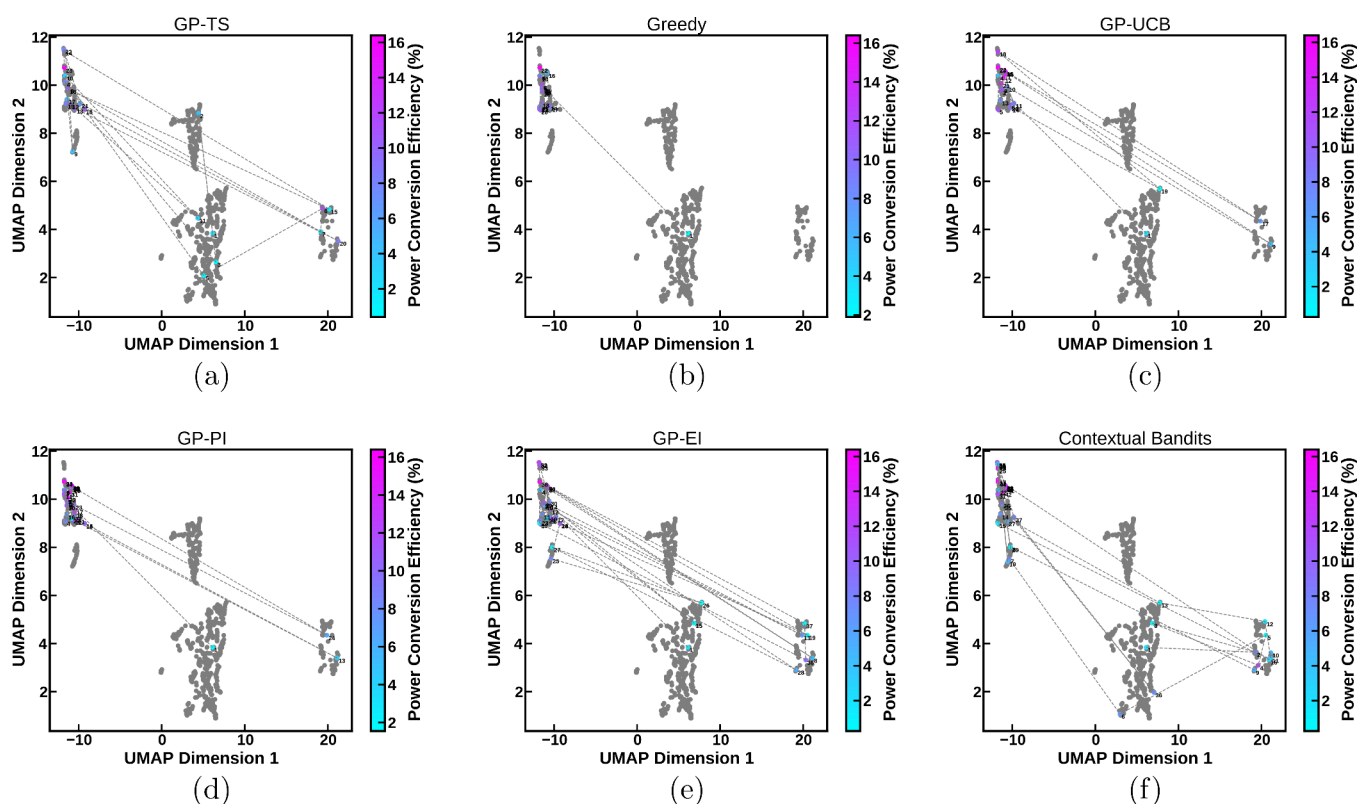
Figure 6. Simulated path of material systems for each data selection method with the material system vectors being embedded in two dimensions using UMAP. The left most cluster in each figure corresponds to nonfullerene small molecule acceptors. The middle two clusters correspond to fullerene acceptors within which the top cluster corresponds to donors with fused aromatic rings and the bottom cluster corresponds to all other donors. The rightmost cluster consists of polymer acceptors. The plots correspond to (a) GP-TS, (b) Greedy, (c) GP-UCB, (d) GP-PI, (e) GP-EI, and (f) linear contextual bandits.

are involved in this data-driven workflow. It may even be possible for funding agencies to allocate resources more efficiently if the iterative methodology described in this section were to be followed. (3) Our simplified problem formulation considers only the discovery of donor/acceptor pairs which are the most critical components of a polymer solar cell and does not consider processing/fabrication.

As we see in Figure 5d−f, the data selection methods used in this work outperform trial and error and random data selection. We compared different methods using the discovery time speedup factor (DTSF) which is defined below:

$$\mathrm{DTSF} = \frac{t_{\max}^{\exp} - t_0}{t_{\max}^{\sim} - t_0} \tag{2}$$

where t_0 is the timestamp associated with the earliest data point in PolymerSolarCells$_{\mathrm{Curated}}$, $t_{\max}^{\exp}$ is the timestamp for the paper when the highest experimental PCE was reported in PolymerSolarCells$_{\mathrm{Curated}}$, $t_{\max}^{\sim}$ is the timestamp corresponding to the index at which the simulated path reaches the same maximum PCE. The timestamp here is the date converted to a rational number in which the fractional part is the day and month converted to a fraction of a full year and the integer part is the year.

Intuitively this gives a sense of how much faster machine learning methods can reach the same end-point as trial and error methods. When fullerenes, as well as nonfullerenes, are used as candidate acceptors, we obtain the greatest speedup of a factor of about 4 while the speedup factor is lower when only one of fullerenes or nonfullerenes are used as candidate

acceptors, reducing to a factor of ∼2. In the context of polymer solar cell donor/acceptor discovery, the former translates to a time saving of ∼15 years. This is one of the first quantitative estimates of the time saved by using machine learning for materials discovery that we are aware of using historical data of how materials discovery evolved. Figure 5a−c assume a fixed starting donor/acceptor system, namely, OC$_1$C$_{10}$-PPV/PC$_{61}$BM. For all methods tested except GP-TS and linear contextual bandits, a fixed starting point would result in a fixed path of donor/acceptor pairs being selected as a GPR model will always generate the same predictions given fixed training data and fixed hyperparameters. In the case of GP-TS and linear contextual bandits, however, there is a sampling step involved in picking a new material which inherently introduces randomness. We used the first ten donor/acceptor pairs reported in PolymerSolarCells$_{\mathrm{Curated}}$ to estimate the range of speedups obtained for each method. These ten donor/acceptor systems belong to the same generation of polymer solar cells and are listed in Supporting Information, Section 3. When error bars are accounted for, we notice, from Figure 5d−f that GP-UCB, GP-EI, GP-PI, and contextual bandits perform similarly. GP-TS appears to underperform the other data selection methods in Figure 5d and f while Greedy outperforms the other data selection methods in Figure 5f. Since Greedy has a higher DTSF on average, we plot the path generated by Greedy in Figure 5a−c. GP-UCB has the tightest variance across all data selection methods tested, for all acceptor types. This indicates that trading off exploration and exploitation is a sound strategy to be robust to the starting
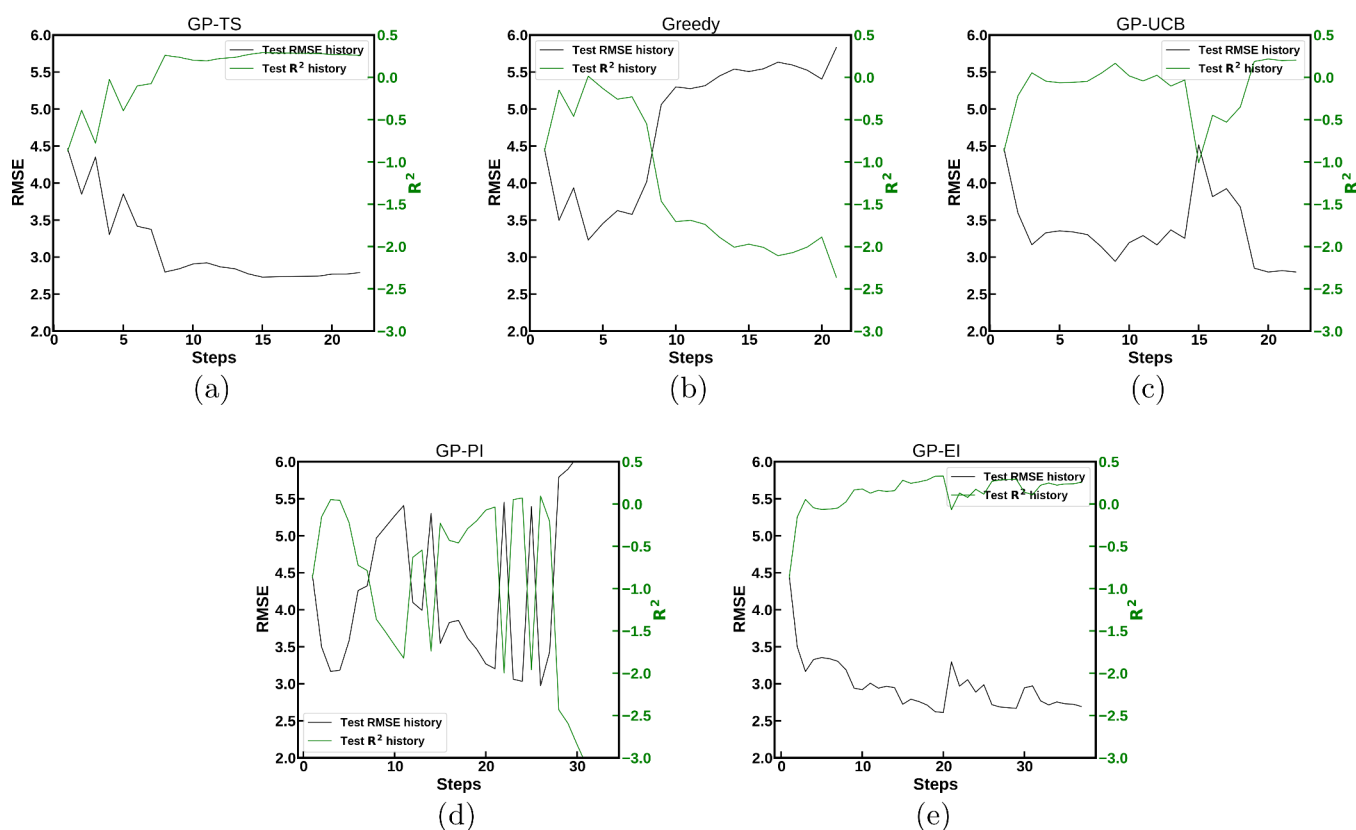
Figure 7. Evolution of the predictive performance of the models trained using different data selection methods using OC_1C_{10}-PPV/PC_{61}BM as the initial donor/acceptor pair. The data selection methods compared are (a) GP-TS, (b) Greedy, (c) GP-UCB, (d) GP-PI, and (e) GP-EI. Each step on the X-axis corresponds to an iteration of the active learning cycle.

material system of the active learning cycle. Note that linear contextual bandits, which is a linear method and therefore much faster than nonparametric methods like GPR, has a similar speedup factor as GPR-based methods. This indicates that the predictive capabilities of a model have little to do with how quickly it can identify the best-performing material system. We shall explore this more fully in the next section.

Analyzing the Predictions from Data Selection Methods. This section examines the path of donor/acceptor pairs picked by various data selection strategies. We projected the vector representations of the concatenated donor/acceptor material fingerprint to two dimensions using Uniform Manifold Approximation and Projection (UMAP).[73] We considered fullerene as well as nonfullerene acceptors. This projection results in four clusters (Figure 6). The leftmost cluster corresponds to nonfullerene small molecule acceptors, and the remaining clusters are described in the caption. The paths shown in Figure 6 were generated using OC_1C_{10}-PPV/ PC_{61}BM as the initial donor/acceptor pair. Note from Figure 6b that Greedy jumps to the nonfullerene small molecule space after the first iteration and stays there during all subsequent iterations. This space is known to contain the candidates most likely to have high power conversion efficiency. This is likely why Greedy has higher average speedups compared to other methods as seen in Figure 5a and c. All other methods spend at least one or more iterations exploring other parts of the donor/ acceptor space. Observe also that Greedy when applied to nonfullerene acceptors and both fullerene and nonfullerene acceptors differs in the first material system only. All subsequent iterations for both cases are restricted to the

space of nonfullerenes. Yet, there is a statistically significant difference in the time taken to find the optimal material system in both cases (refer Figure 5d and f) with the nonfullerene only case taking significantly longer. We posit that the initial inclusion of the fullerene acceptor or more generally the inclusion of some initial material system that is different from the space where the optimal material system is expected to lie would accelerate convergence to the optimal material system. This could be because the initial diversity helps give the model a better "view" of the material space compared to if the initial candidates were less diverse. The latter can at best present the model with a very local region of the material space, thus requiring more iterations to understand the best direction to traverse in material space.

At each iteration, we measured the predictive performance of the trained GPR model by comparing its predictions against all points not yet added to the test set (Figure 7). The data points used for measuring test performance are reduced by one point at every iteration and have some points that are different for each selection method. However, the number of points selected during a typical active learning run (20−30) is small compared to the total number of data points (835) which makes this a reasonable approach to infer the general trend in the predictive capabilities of the models being trained. There is no consistent improvement in the predictive performance of the models being trained as the number of data points in the training set increases, for most of the selection methods we test. This is consistent with results reported in ref 14. The notable exception to this however is GP-TS in which successive models have improved predictive performance.

This intuitively makes sense as all other methods pick material systems by directly optimizing for high power conversion efficiency and therefore bias the trained model using the picked data points. GP-TS on the other hand picks material systems by maximizing over PCEs sampled from the predicted distribution for each material system and thus samples a more diverse set of data points as seen in Figure 6a. This diverse sampling is likely what leads to GP-TS being slower than other data selection methods as seen in Figure 5d−f. Thus, this data selection method allows trading-off speed of discovery for training a stronger predictive model. By design, GP-TS reaches the optimal material state quickly and in parallel trains models with strong predictive performance and is the only method out of the methods we test that consistently achieves this balance across acceptor types. The trend observed in Figure 7 is reported for both types of acceptors. The same trend for GP-TS is observed for iterations performed over fullerene acceptors and nonfullerene acceptor candidates, reported in Supporting Information, Section 4.

■ SUMMARY AND OUTLOOK

A pipeline that extracts material property data from published literature which is then used for training property prediction models was built. We found promising donor/acceptor combinations that have not yet been reported in the literature, by training models to predict power conversion efficiency. We demonstrate that data-driven methods alone would have discovered the most promising donor/acceptor system in only one-fourth of the time it took through trial and error. This allows for a stronger case to be made to policymakers to encourage further use of ML methods for materials discovery. We summarize some best practices learned from this study for deploying active learning for materials discovery: (1) Gaussian Process-Thompson sampling should be used as the active learning strategy if training a strong property predictor is desired. (2) If robustness to the starting material is desired, then GP-UCB is the recommended strategy which will also on average lead to fast discovery of the optimal material system. (3) Using a diverse initial set of candidates during the active learning cycle helps the model learn the optimization landscape more quickly and is empirically observed to lead to faster convergence.

We highlight some key limitations of our study: (1) The list of candidate material systems we pick from is the set of donor/acceptor systems that have already been tested in the literature. The structures that were tested much later in the literature could only be discovered due to the experimental trial and error that preceded it. (2) We assume that the time intervals of measurement in the data-driven view are identical to the trial and error approach, i.e., new materials through data-driven approaches can be discovered at the same rate at which experimental papers containing new donor/acceptor systems were published. (3) Our simplified formulation of the polymer solar cell optimization problem does not consider other aspects of the cell such as the electrodes, electron/hole transport layer, and fabrication/processing of the cell. These factors would affect our estimate of the speedup of ML methods over trial-and-error experiments.

This study views active learning as a sequential process whereby a material is "chosen" by the model and tested by the scientist. The only input from the scientist is the list of materials provided initially. The difficulty of coming up with a perfect initial list ensures, however, that scientists and machines would need to coordinate more closely in practice. The insights generated by the selection process can guide scientific intuition to update the list of materials continuously. An example of this is provided in Supporting Information, Section 5.

One of the bottlenecks when going from NLP-extracted data to a data set usable for training ML models is obtaining SMILES strings for the donor and acceptor. This had to be done manually during data curation. There are computer vision tools available that can convert structures to SMILES strings such as MolScribe,[74] Decimer,[75] and OSRA.[76] However, donor or acceptor structures found in the literature are typically part of a larger figure such as a reaction scheme. Robust segmentation of relevant structures combined with conversion of structures to SMILES strings is necessary to bridge the gap between NLP extracted data and trained property predictors. Furthermore, the conversion of polymer structures to SMILES strings is not yet possible through these tools.

We also limited our focus to optimizing a single material property, namely the PCE. In practice, however, materials scientists are often interested in optimizing multiple properties simultaneously which may be inversely corelated. In ref 10, for instance, the authors find optimal multiprincipal element alloys while optimizing two ductility indicators (Pugh's Ratio and Cauchy pressure). Multiobjective Bayesian approaches using hypervolume indicators are promising approaches for solving such problems.[77,78] In organic photovoltaics, a possible direction of inquiry would be a simultaneous optimization of the efficiency, flexibility, and stability of the solar cell. PSCs need to be stable and retain their power conversion efficiency with usage. This is typically inversely corelated with PCE. Higher flexibility PSCs usually have weaker mechanical properties and therefore lower PCE.[79−81]

The ability to scale up the pipeline we have built to other properties and materials classes will significantly speed up the development and deployment of new materials. This in turn will accelerate scientific discovery across many critical applications such as clean energy, healthcare, and devices.

■ ASSOCIATED CONTENT

Data Availability Statement

All code and data needed to reproduce the results in this paper can be found at https://github.com/pranav-s/PolymerSolarCellsML.

⬡ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.4c00709.

> Linear contextual bandits, fingerprinting methodology, initial donor/acceptor systems used, performance of trained models at each iteration over test set, and analyzing structures of donors and acceptors (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Rampi Ramprasad − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* ⬡ orcid.org/0000-0003-4630-1565; Email: rampi.ramprasad@mse.gatech.edu

Authors

Pranav Shetty − *School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* ⬤ orcid.org/0000-0003-2015-9556

Aishat Adeboye − *School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Sonakshi Gupta − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Chao Zhang − *School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.chemmater.4c00709

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785−796.

(2) Ma, R.; Sharma, V.; Baldwin, A. F.; Tefferi, M.; Offenbach, I.; Cakmak, M.; Weiss, R.; Cao, Y.; Ramprasad, R.; Sotzing, G. A. Rational design and synthesis of polythioureas as capacitor dielectrics. *Journal of Materials Chemistry A* **2015**, *3*, 14845−14852.

(3) Wu, C.; Chen, L.; Deshmukh, A.; Kamal, D.; Li, Z.; Shetty, P.; Zhou, J.; Sahu, H.; Tran, H.; Sotzing, G.; Ramprasad, R.; Cao, Y. Dielectric polymers tolerant to electric field and temperature extremes: Integration of phenomenology, informatics, and experimental validation. *ACS Appl. Mater. Interfaces* **2021**, *13*, 53416−53424.

(4) He, J.; Rabe, K. M.; Wolverton, C. Computationally accelerated discovery of functional and structural Heusler materials. *MRS Bull.* **2022**, *47*, 559−572.

(5) Jia, X.; Deng, Y.; Bao, X.; Yao, H.; Li, S.; Li, Z.; Chen, C.; Wang, X.; Mao, J.; Cao, F.; Sui, J.; Wu, J.; Wang, C.; Zhang, Q.; Liu, X. Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials. *npj Computational Materials* **2022**, *8*, 34.

(6) Kranthiraja, K.; Saeki, A. Machine Learning-Assisted Polymer Design for Improving the Performance of Non-Fullerene Organic Solar Cells. *ACS Appl. Mater. Interfaces* **2022**, *14*, 28936−28944.

(7) 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. *Science Advances* **2019**, *5*, No. eaay4275.

(8) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* **2022**, *8*, No. eabn9545.

(9) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott,

M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* **2020**, *6*, No. eaaz8867.

(10) Khatamsaz, D.; Vela, B.; Singh, P.; Johnson, D. D.; Allaire, D.; Arróyave, R. Multi-objective materials bayesian optimization with active learning of design constraints: Design of ductile refractory multi-principal-element alloys. *Acta Mater.* **2022**, *236*, 118133.

(11) Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Communications* **2019**, *9*, 860−866.

(12) Rohr, B.; Stein, H. S.; Guevarra, D.; Wang, Y.; Haber, J. A.; Aykol, M.; Suram, S. K.; Gregoire, J. M. Benchmarking the acceleration of materials discovery by sequential learning. *Chemical science* **2020**, *11*, 2696−2706.

(13) Wang, A.; Liang, H.; McDannald, A.; Takeuchi, I.; Kusne, A. G. Benchmarking active learning strategies for materials optimization and discovery. *Oxford Open Materials Science* **2022**, *2*, itac006.

(14) Borg, C. K.; Muckley, E. S.; Nyby, C.; Saal, J. E.; Ward, L.; Mehta, A.; Meredig, B. Quantifying the performance of machine learning models in materials discovery. *Digital Discovery* **2023**, *2*, 327−338.

(15) Duros, V.; Grizou, J.; Xuan, W.; Hosni, Z.; Long, D.-L.; Miras, H. N.; Cronin, L. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angew. Chem.* **2017**, *129*, 10955−10960.

(16) Duros, V.; Grizou, J.; Sharma, A.; Mehr, S. H. M.; Bubliauskas, A.; Frei, P.; Miras, H. N.; Cronin, L. Intuition-Enabled Machine Learning Beats the Competition When Joint Human-Robot Teams Perform Inorganic Chemical Experiments. *J. Chem. Inf. Model.* **2019**, *59*, 2664−2671.

(17) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89−96.

(18) Fu, H.; Wang, Z.; Sun, Y. Polymer donors for high-performance non-fullerene organic solar cells. *Angew. Chem. Int.* **2019**, *58*, 4442−4453.

(19) Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *J. Phys. Chem. Lett.* **2018**, *9*, 2639−2646.

(20) Miyake, Y.; Saeki, A. Machine learning-assisted development of organic solar cell materials: issues, analyses, and outlooks. *J. Phys. Chem. Lett.* **2021**, *12*, 12391−12401.

(21) Greenstein, B. L.; Hutchison, G. R. Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms. *J. Phys. Chem. C* **2023**, *127*, 6179−6191.

(22) Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A. The Harvard organic photovoltaic dataset. *Sci. Data* **2016**, *3*, 1−7.

(23) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95−98.

(24) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **2019**, *6*, 1−11.

(25) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Automated Chemical Reaction Extraction from Scientific Literature. *J. Chem. Inf. Model.* **2022**, *62*, 2035−2045.

(26) Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **2019**, *59*, 3692−3702.

(27) Court, C. J.; Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **2018**, *5*, 1−12.

(28) Jensen, Z.; Kim, E.; Kwon, S.; Gani, T. Z.; Roman-Leshkov, Y.; Moliner, M.; Corma, A.; Olivetti, E. A machine learning approach to

zeolite synthesis enabled by automatic literature data extraction. *ACS central science* **2019**, *5*, 892−899.

(29) Cheung, J.; Zhuang, Y.; Li, Y.; Shetty, P.; Zhao, W.; Grampurohit, S.; Ramprasad, R.; Zhang, C. *POLYIE: A Dataset of Information Extraction from Polymer Material Scientific Literature.* Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). Mexico City, Mexico, 2024; pp 2370−2385, DOI: 10.18653/v1/2024.naacl-long.131.

(30) Hiszpanski, A. M.; Gallagher, B.; Chellappan, K.; Li, P.; Liu, S.; Kim, H.; Han, J.; Kailkhura, B.; Buttler, D. J.; Han, T. Y.-J. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* **2020**, *60*, 2876−2887.

(31) Court, C. J.; Jain, A.; Cole, J. M. Inverse Design of Materials That Exhibit the Magnetocaloric Effect by Text-Mining of the Scientific Literature and Generative Deep Learning. *Chem. Mater.* **2021**, *33*, 7217−7231.

(32) Tchoua, R. B.; Ajith, A.; Hong, Z.; Ward, L. T.; Chard, K.; Belikov, A.; Audus, D. J.; Patel, S.; de Pablo, J. J.; Foster, I. T. Creating training data for scientific named entity recognition with minimal human effort. *International Conference on Computational Science.* **2019**, *11536*, 398−411.

(33) Shetty, P.; Ramprasad, R. Automated knowledge extraction from polymer literature using natural language processing. *iScience* **2021**, *24*, 101922.

(34) Shetty, P.; Ramprasad, R. Machine-Guided Polymer Knowledge Extraction Using Natural Language Processing: The Example of Named Entity Normalization. *J. Chem. Inf. Model.* **2021**, *61*, 5377−5385.

(35) Shetty, P.; Rajan, A. C.; Kuenneth, C.; Gupta, S.; Panchumarti, L. P.; Holm, L.; Zhang, C.; Ramprasad, R. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials* **2023**, *9*, 52.

(36) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(37) Lee, S.; Jeong, D.; Kim, C.; Lee, C.; Kang, H.; Woo, H. Y.; Kim, B. J. Eco-friendly polymer solar cells: Advances in green-solvent processing and material design. *ACS Nano* **2020**, *14*, 14493−14527.

(38) Sorrentino, R.; Kozma, E.; Luzzati, S.; Po, R. Interlayers for non-fullerene based polymer solar cells: distinctive features and challenges. *Energy Environ. Sci.* **2021**, *14*, 180−223.

(39) Jin, J.; Wang, Q.; Ma, K.; Shen, W.; Belfiore, L. A.; Bao, X.; Tang, J. Recent developments of polymer solar cells with photovoltaic performance over 17. *Adv. Funct. Mater.* **2023**, *33*, 2213324.

(40) Yin, H.; Yan, C.; Hu, H.; Ho, J. K. W.; Zhan, X.; Li, G.; So, S. K. Recent progress of all-polymer solar cells−From chemical structure and device physics to photovoltaic performance. *Materials Science and Engineering: R: Reports* **2020**, *140*, 100542.

(41) Ma, S.; Zhang, H.; Feng, K.; Guo, X. Polymer Acceptors for High-Performance All-Polymer Solar Cells. *Chemistry−A European Journal* **2022**, *28*, No. e202200222.

(42) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14*, 4099.

(43) Zhao, J.; Yao, C.; Ali, M. U.; Miao, J.; Meng, H. Recent advances in high-performance organic solar cells enabled by acceptor−donor−acceptor−donor−acceptor (A−DA' D−A) type acceptors. *Materials Chemistry Frontiers* **2020**, *4*, 3487−3504.

(44) Padula, D.; Troisi, A. Concurrent optimization of organic donor−acceptor pairs through machine learning. *Adv. Energy Mater.* **2019**, *9*, 1902463.

(45) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, 171104.

(46) Gurnani, R.; Kuenneth, C.; Toland, A.; Ramprasad, R. Polymer Informatics at Scale with Multitask Graph Neural Networks. *Chem. Mater.* **2023**, *35*, 1560−1567.

(47) Kuenneth, C.; Schertzer, W.; Ramprasad, R. Copolymer informatics with multitask deep neural networks. *Macromolecules* **2021**, *54*, 5957−5961.

(48) Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington, T.; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **2018**, *148*, 241702 DOI: 10.1063/1.5003074.

(49) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.

(50) Williams, C. K.; Rasmussen, C. E. *Gaussian processes for machine learning*; MIT Press: Cambridge, MA, 2006; Vol. 2.

(51) Noack, M. M.; Doerk, G. S.; Li, R.; Streit, J. K.; Vaia, R. A.; Yager, K. G.; Fukuto, M. Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. *Sci. Rep.* **2020**, *10*, 17663.

(52) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **2019**, *5*, 21.

(53) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology* **2020**, *1*, 025006.

(54) Srinivas, N.; Krause, A.; Kakade, S. M.; Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

(55) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization* **1998**, *13*, 455−492.

(56) Agrawal, S.; Goyal, N. Thompson sampling for contextual bandits with linear payoffs. *International Conference on Machine Learning*, 2013; pp 127−135.

(57) Bin, H.; van der Pol, T. P.; Li, J.; van Gorkom, B. T.; Wienk, M. M.; Janssen, R. A. Efficient organic solar cells with small energy losses based on a wide-bandgap trialkylsilyl-substituted donor polymer and a non-fullerene acceptor. *Chemical Engineering Journal* **2022**, *435*, 134878.

(58) Jain, S.; van Zuylen, M.; Hajishirzi, H.; Beltagy, I. SciREX: A challenge dataset for document-level information extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, online, 2020; DOI: 10.18653/v1/2020.acl-main.670.

(59) Song, C. E.; Ham, H.; Noh, J.; Lee, S. K.; Kang, I.-N. Efficiency enhancement of a fluorinated wide-bandgap polymer for ternary nonfullerene organic solar cells. *Polymer* **2020**, *188*, 122131.

(60) Wang, H.; Liu, T.; Zhou, J.; Mo, D.; Han, L.; Lai, H.; Chen, H.; Zheng, N.; Zhu, Y.; Xie, Z.; He, F. Bromination: an alternative strategy for non-fullerene small molecule acceptors. *Advanced Science* **2020**, *7*, 1903784.

(61) He, Q.; Shahid, M.; Wu, J.; Jiao, X.; Eisner, F. D.; Hodsden, T.; Fei, Z.; Anthopoulos, T. D.; McNeill, C. R.; Durrant, J. R.; Heeney, M. Fused cyclopentadithienothiophene acceptor enables ultrahigh short-circuit current and high efficiency > 11% in as-cast organic solar cells. *Adv. Funct. Mater.* **2019**, *29*, 1904956.

(62) Li, G.; Cheng, H.; Zhang, Y.; Yang, T.; Liu, Y. Higher open circuit voltage caused by chlorinated polymers endows improved efficiency of binary organic solar cell. *Org. Electron.* **2020**, *83*, 105776.

(63) Huang, J.; Peng, R.; Xie, L.; Song, W.; Hong, L.; Chen, S.; Wei, Q.; Ge, Z. A novel polymer donor based on dithieno [2, 3-d: 2', 3'-d"] benzo [1, 2-b: 4, 5-b'] dithiophene for highly efficient polymer solar cells. *Journal of Materials Chemistry A* **2019**, *7*, 2646−2652.

(64) Firdaus, Y.; Maffei, L. P.; Cruciani, F.; Müller, M. A.; Liu, S.; Lopatin, S.; Wehbe, N.; Ndjawa, G. O. N.; Amassian, A.; Laquai, F.; Beaujuge, P. M. Polymer Main-Chain Substitution Effects on the

Efficiency of Nonfullerene BHJ Solar Cells. *Adv. Energy Mater.* **2017**, *7*, 1700834.

(65) Zhu, L.; Zhong, W.; Qiu, C.; Lyu, B.; Zhou, Z.; Zhang, M.; Song, J.; Xu, J.; Wang, J.; Ali, J.; Feng, W.; Shi, Z.; Gu, X.; Ying, L.; Zhang, Y.; Liu, F. Aggregation-induced multilength scaled morphology enabling 11.76% efficiency in all-polymer solar cells using printing fabrication. *Adv. Mater.* **2019**, *31*, 1902899.

(66) Li, X.; Weng, K.; Ryu, H. S.; Guo, J.; Zhang, X.; Xia, T.; Fu, H.; Wei, D.; Min, J.; Zhang, Y.; Woo, H. Y.; Sun, Y. Non-Fullerene Organic Solar Cells Based on Benzo [1, 2-b: 4, 5-b'] difuran-Conjugated Polymer with 14% Efficiency. *Adv. Funct. Mater.* **2020**, *30*, 1906809.

(67) Ma, L.; Zhang, S.; Yao, H.; Xu, Y.; Wang, J.; Zu, Y.; Hou, J. High-efficiency nonfullerene organic solar cells enabled by 1000 nm thick active layers with a low trap-state density. *ACS Appl. Mater. Interfaces* **2020**, *12*, 18777−18784.

(68) Sun, C.; Pan, F.; Chen, S.; Wang, R.; Sun, R.; Shang, Z.; Qiu, B.; Min, J.; Lv, M.; Meng, L.; Zhang, C.; Xiao, M.; Yang, C.; Li, Y. Achieving fast charge separation and low nonradiative recombination loss by rational fluorination for high-efficiency polymer solar cells. *Adv. Mater.* **2019**, *31*, 1905480.

(69) Zhang, C.; Song, X.; Liu, K.-K.; Zhang, M.; Qu, J.; Yang, C.; Yuan, G.-Z.; Mahmood, A.; Liu, F.; He, F.; Baran, D.; Wang, J.-L. Electron-Deficient and Quinoid Central Unit Engineering for Unfused Ring-Based A1−D−A2−D−A1-Type Acceptor Enables High Performance Nonfullerene Polymer Solar Cells with High Voc and PCE Simultaneously. *Small* **2020**, *16*, 1907681.

(70) He, E.; Zheng, Z.; Lu, Y.; Guo, F.; Gao, S.; Pang, X.; Mola, G. T.; Zhao, L.; Zhang, Y. Highly efficient non-fullerene polymer solar cells from a benzo [1, 2-b: 4, 5-b'] difuran-based conjugated polymer with improved stabilities. *Journal of Materials Chemistry A* **2020**, *8*, 11381−11390.

(71) Fujinuma, N.; DeCost, B.; Hattrick-Simpers, J.; Lofland, S. E. Why big data and compute are not necessarily the path to big materials science. *Communications Materials* **2022**, *3*, 59.

(72) Dyakonov, V. Electrical aspects of operation of polymer−fullerene solar cells. *Thin Solid Films* **2004**, *451*, 493−497.

(73) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, *1802*, 03426.

(74) Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; Barzilay, R. MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. *J. Chem. Inf. Model.* **2023**, *63*, 1925−1934.

(75) Rajan, K.; Zielesny, A.; Steinbeck, C. DECIMER: towards deep learning for chemical image recognition. *J. Cheminformatics* **2020**, *12*, 1−9.

(76) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49*, 740−743.

(77) Yang, Q.; Ding, S. Novel algorithm to calculate hypervolume indicator of Pareto approximation set. *International Conference on Intelligent Computing*; ICIC, 2007; pp 235−244.

(78) Emmerich, M. T.; Deutz, A. H.; Klinkenberg, J. W. *Hypervolume-based expected improvement: Monotonicity properties and exact computation*; 2011 IEEE Congress of Evolutionary Computation (CEC), 2011; pp 2147−2154.

(79) Xu, X.; Fukuda, K.; Karki, A.; Park, S.; Kimura, H.; Jinno, H.; Watanabe, N.; Yamamoto, S.; Shimomura, S.; Kitazawa, D.; Yokota, T.; Umezu, S.; Nguyen, T.-Q.; Someya, T. Thermally stable, highly efficient, ultraflexible organic photovoltaics. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 4589−4594.

(80) Gevorgyan, S. A.; Heckler, I. M.; Bundgaard, E.; Corazza, M.; Hösel, M.; Søndergaard, R. R.; dos Reis Benatto, G. A.; Jørgensen, M.; Krebs, F. C. Improving, characterizing and predicting the lifetime of organic photovoltaics. *J. Phys. D: Appl. Phys.* **2017**, *50*, 103001.

(81) Rafique, S.; Abdullah, S. M.; Sulaiman, K.; Iwamoto, M. Fundamentals of bulk heterojunction organic solar cells: An overview of stability/degradation issues and strategies for improvement. *Renewable and Sustainable Energy Reviews* **2018**, *84*, 43−53.