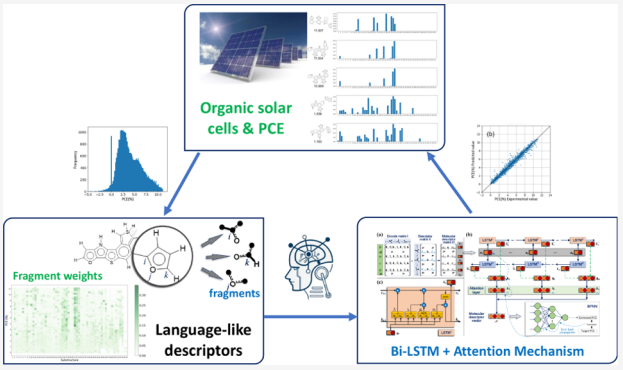
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR
Article

Deep-Learning Architecture in QSPR Modeling for the Prediction of Energy Conversion Efficiency of Solar Cells

Jinkui Wu, Shihui Wang, Li Zhou,* Xu Ji, Yiyang Dai, Yagu Dang, and Markus Kraft

Cite This: *Ind. Eng. Chem. Res.* 2020, 59, 18991−19000

ABSTRACT: The efficient and effective design of chemical processes and products heavily relies on the accurate prediction of essential properties. In this work, a deep-learning architecture integrating a bidirectional long short-term memory (Bi-LSTM) network, an attention mechanism, and a back-propagation neural network (BPNN) is developed for the prediction of energy conversion efficiency of organic solar cells. Inspired by the success of artificial intelligence in natural language processing, we first developed a novel strategy for molecular signature encoding and information embedding in order to depict the compositional structures of molecules. Then, an advanced recurrent neural network named Bi-LSTM is employed to process the molecular information, while the BPNN is applied to correlate energy conversion efficiency. During this procedure, the attention mechanism is used to identify molecular constituents that are important to the property of interest. To evaluate the performance of the proposed approach, the energy conversion efficiencies of more than 20,000 organic photovoltaics are used to train and test the model. Result comparisons with several other modeling approaches indicate that the proposed method is competitive in prediction accuracy and possesses good transferability to small data sets. Additionally, the proposed method is capable of identifying decisive molecular constituents, providing instructive information for the reverse design of organic solar cells.



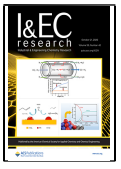
INTRODUCTION

Solar energy harvesting based on organic photovoltaics (OPVs) is a promising sustainable way of addressing growing global energy needs while minimizing detrimental environmental effects. The development of OPVs is still challenging and heavily relies on the improvement of power conversion efficiency (PCE). The community strives to tackle this problem from two perspectives, the development of advanced synthesis processes[1] and the discovery of novel compounds.[2] It is acknowledged that the achievable performance of a material is restrained by its microstructure morphology. Therefore, the identification of more suitable chemical compounds is essential. Traditionally, the discovery of new compounds for application is dominated by experiment-driven trial-and-error methods, which are costly (in resources and time) and the effectiveness (in exploring the enormous chemical space) is limited.

Along with the rapid development in computing power and successful stories in artificial intelligence, the development of prediction methods to accelerate process and product design has gained momentum in various fields.[3−6] In the field of OPV, the Scharber model[7] is widely used for estimating maximum PCE of bulk-heterojunction solar cells from the energy levels of the lowest unoccupied molecular orbital (LUMO) of the acceptor and the highest occupied molecular orbital (HOMO) of the donor. Although the models are usually too simplistic to account for all the complicated physicochemical behavior of an organic solar cell, it can provide a valuable indication about the potential performance that a candidate compound may have achieved. Recently, it was reported that a new peak of PCE (17.35%) has been achieved by following the guidance of a semiempirical model analysis.[8]

It is ascertained that models reflecting the relationships between the quantitative structure and property of compounds (a.k.a. QSPR model) can provide valuable support for the directed improvement of material properties. In recent years, high-throughput virtual screening (HTVS) is becoming popular in the discovery of novel high-performance materials.[9−12] Centered around the HTVS, many research efforts have been carried out worldwide, including the establishment of theoretical/experimental property database of compounds,

Received: August 6, 2020
Revised: October 1, 2020
Accepted: October 2, 2020
Published: October 13, 2020



https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

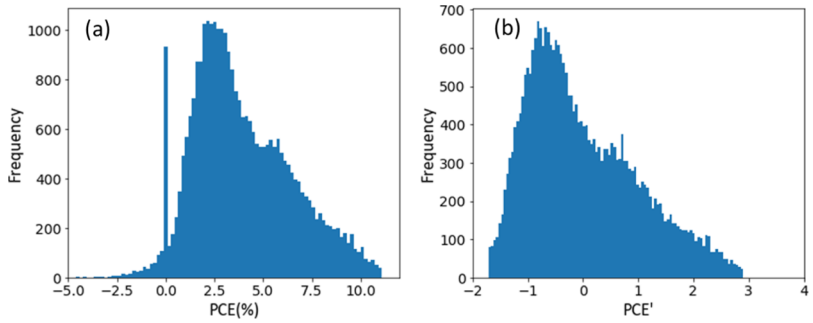


Figure 1. Distribution of data in the CEPDB: (a) before data preprocessing and (b) after data preprocessing.

the development of the forward property prediction model, and reverse material design rules and algorithms.

The University of Harvard has established three authoritative databases for organic solar cells and contributed to the community as open source resources, including the Harvard Clean Energy Project Database (CEPDB),[13] the Harvard Organic Photovoltaic Dataset (HOPV15),[14] and the nonfullerene small-molecule acceptors database (NFADB).[15] The CEPDB contains energies of the HOMO/LUMO of 2.3 million organic compounds and their calculated PCE values based on the Scharber model.[7] The NFADB is a collection of 50,000 nonfullerene acceptor materials with calculated PCE values from calibrated HOMO/LUMO energies.[15] HOPV15 is a small collation of experimental photovoltaic data from the literature and the corresponding quantum-chemical calculations performed over a range of conformers.[14] Although the data were obtained based on various assumptions and are subject to limitations, insights and design principles can be derived from the phenomena that the data reveal. Hachmann et al.[16] identified critical building blocks and extracted design rules for potential OPV candidates by analyzing data from the CEPDB.

As for the development of forward property prediction models, it usually consists of two parts, encoding and mapping. The encoding process converts the chemical constitution and structure information into a series of numeric values called descriptors, while the mapping process determines a desirable function to map between the descriptors and the property of interest. Padula et al.[17] proposed to combine electronic and structural features as descriptors for the property prediction of organic solar cells, based on which the predictive capability of the machine learning model reached a correlation of $r \approx 0.7$. Sahu et al.[18] used 13 microscopic properties obtained from quantum chemical calculations as descriptors for the PCE prediction of small organic molecules, and a gradient boosting model with Pearsons coefficient equal to 0.79 was reported. Later on, the same research group furthered the study by introducing several new microscopic attributes, such as ground-state geometries and cations and anions, as descriptors, and the achieved model correlation is 0.78.[19] Furthermore, Sun et al.[20] applied convolution to extract descriptors from the pictures of the molecular structure for the PCE performance estimation of organic compounds, and a convolutional neural network model with an accuracy of 91.2% was achieved. With the development of machine learning, some state-of-the-art strategies such as transfer learning have also been employed to improve agreement between predictions and experiments.[21]

Although many efforts have been made, challenges still remain for the future development of OPVs. First, descriptors generated by quantum chemical calculation can deliver accurate results, but they usually require in-depth domain knowledge from the modeler and are computationally expensive, which limits their effectiveness in HTVS. Second, the micro-physicochemical environment of a compound is very complex. It is easy for the easily accessible descriptors to leave out important chemical information, which leads to less satisfactory prediction results, and many machine learning models are "black boxes", whose results are usually of low interpretability. This work strives to alleviate this dilemma.

This work proposes a novel language-like molecular extraction and representing strategy and based on which a deep-learning model containing attention mechanism for the prediction of PCE values of organic compounds, striving to facilitate virtual screening of high-performance OPVs and offering some possible explanations in the prediction of deep learning. It is inspired by the work of Cadeddu et al.,[22] in which the authors verified the similarity between organic chemistry and natural language in terms of the same structure in the frequency of molecular fragments and text fragments, respectively. First, the molecular graph of each organic compound is disassembled into fragments, and each fragment is numbered according to their relative position and connectivity. By denoting each involved fragment as a unique fragment vector, an organic compound can then be represented as a language-like descriptor that takes into account the fragment sequence information. Second, a natural language-processing technique is applied to "comprehend" the descriptor and to correlate the molecular information to its potential PCE performance. Third, the widely adopted neural network interpreter attention mechanism is applied to identify the molecular fragments that are important to the potential PCE performance of organic compounds, aiming to enhance the interpretability of the model and obtain insights for better OPV design.

METHOD

In this section, a deep neural network (DNN) model that integrates the Bi-LSTM network, attention mechanism, and back-propagation neural network (BPNN) is introduced for the determination of the correlation between molecular structures and the PCE values of organic compounds. The process of developing a reliable QSPR model with the proposed method consists of the following four steps.

i Data acquisition and preprocessing. The experimentally measured or quantum-chemical calculations of the PCE values and the SMILES strings of organic compounds are collected[23] (shown in Figure 1).

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

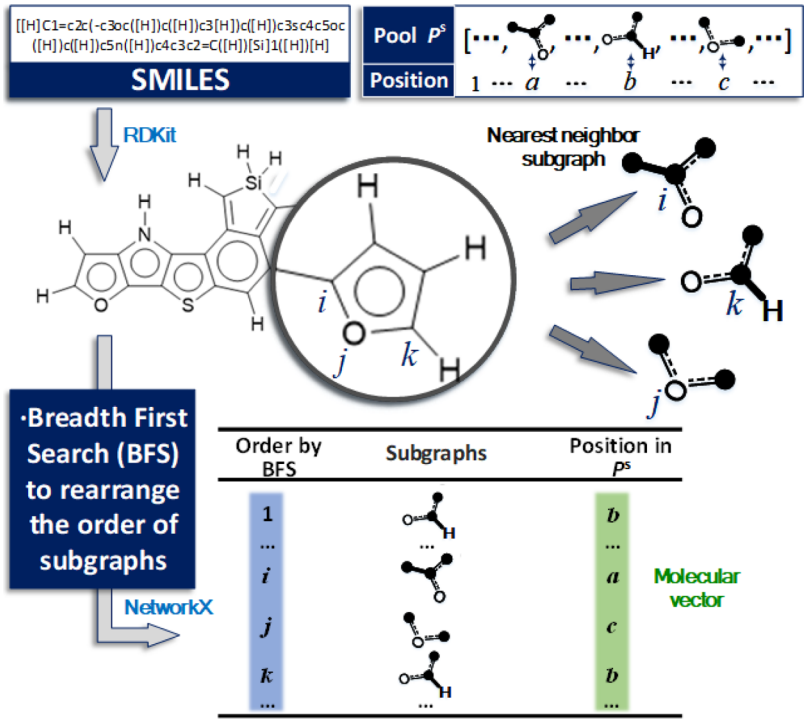


Figure 2. Embedding based on a given molecular graph.

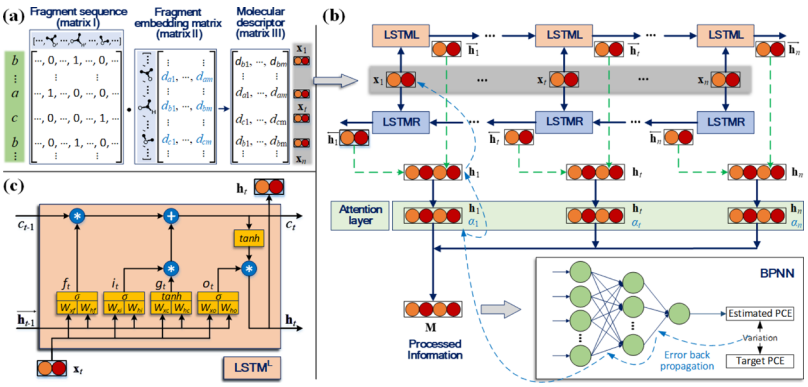


Figure 3. Predictor network: (a) embedding procedure, m is the dimension of the embedded vectors, in this case, m = 128; the encode matrix represents a given molecule. (b) Bi-LSTM network with the attention mechanism and (c) LSTM cell sketch map.

ii Molecular pre-encoding. The SMILES strings of the candidate molecules are used to generate a list of fragment sequences, based on the connectivity of their constructive fragments and the position of each fragment in a predefined fragment pool (shown in Figure 2).

iii Encoder−predictor network training. The fragment sequences are further embedded into molecular descriptors, based on the fragment-embedding matrix, after which the molecular descriptors and the corresponding PCE values are used to train the DNN. In the training process, the descriptor and the DNN model parameters are updated automatically by calculating the gradient of the loss function to improve the model performance in PCE value prediction (shown in Figure 3).

iv Model evaluation. The predictive performance of the developed QSPR model is evaluated by a test data set.

Data Acquisition and Preprocessing. The SMILES strings and PCE values of 25,000 candidate OPVs were collected from the CEPDB.[13] The collected data are processed to eliminate the data items with outliers and missing values, if any, and the unrealistic PCE values, such as negative and zero values, are removed. As a result, 1122 organic compounds are removed from the data set, and the remaining 23,878 molecules are the final data set for employing the deep-learning model. Additionally, Z-score standardization[24] is applied to preprocess the PCE values, so as to accelerate the convergence of the model-training process and improve model accuracy. The formulation of the Z-score standardization is given in eq 1.

$$y_i' = (y_i - \overline{y})/s(y) \tag{1}$$

where y_i is the PCE value for molecular i and $\overline{y}$ and $s(y)$ are the mean value and standard deviation for all the PCE values of the collected OPVs, respectively. Figure 1 gives the distribution of the collected data before and after the data preprocessing.

Pre-encoding of Organic Compounds into Fragment Sequences. It has been validated that organic molecules

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

contain fragments whose rank distribution is essentially identical to that of sentence fragments in natural language.[22] Enlightened by this proven fact and the natural language-processing technology, we propose to disassemble an organic molecule into fragments and arrange the fragments by their connection sequence information. In this way, the organic molecules can be converted into a series of fragment sequences. This process is described in the following.

The SMILES strings are converted into molecular graphs via RDKit and Networkx.[25,26] A molecular graph is a projection of a molecule to the plane, where the vertex represents the atom and the edge is the chemical bond. As shown in Figure 2, to better encode the connectivity information of the fragments, the nearest neighbor subgraph[27] and the Breadth First Search (BFS) algorithm[28] are employed to disassemble the molecule (a graph) into fragments (subgraphs) and arrange the fragments by the respective BFS order. In this way, an ordered fragment sequence is generated. It is worth mentioning that, normally, different organic compounds are composed of different numbers of fragments. In order to make sure that the fragment sequence of the molecules are of the same dimension, the molecule with the most constituent fragments is identified and the count of its constituent fragments is set as the dimension of the fragment sequence, which is 60 in this paper. When generating fragment sequences for other molecules, zero values are padded to assure dimensional consistency.

Herein, the nearest neighbor subgraphs are induced by the nearest neighbor vertices and the edges within a hop from the current vertex, implying that each atom representation, namely a fragment, is the sum of its nearest chemical environment. It is much more effective and efficient than the common way of chemical information embedding that strives to embed all atoms and chemical bonds because there are only a few types of atoms and bonds in OPV molecules. Moreover, Cadeddu et al.[22] also demonstrated that the fragments of organic compounds were similar to the words of sentences. The use of the BFS traversal algorithm takes connectivity information of the constructive fragments of a molecule into consideration, aiming to enhance the QSPR model accuracy. After pre-encoding, each compound is expressed as a unique sequence.

Embedding of the Fragment Sequence into a Language-like Descriptor and Predictor Model Training. The generated fragment sequence is represented as a one-hot matrix (denoted as matrix I), which is shown in Figure 3a. A fragment-embedding matrix (matrix II) that contains the embedding vectors (each row) of all the involved fragments in the pool is generated.The product of matrix I and matrix II results in the molecular descriptor (matrix III) for the corresponding compound. The molecular descriptor is then passed to a Bi-LSTM network. In order to effectively extract the fragment sequence information, each row (a fragment vector) of the descriptor is simultaneously processed by a forward-directional LSTM cell and a backward-directional LSTM cell. The processed fragment information is denoted as $\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_t}$, respectively, which are then fed into the subsequent forward and backward directional LSTM cell.

The combination of $\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_t}$ forms the input of the attention layer.[29] On the attention layer, normalized weights (α_t) are applied to indicate the importance of the fragments and to improve the performance of the model. After the attention layer, the processed information $\mathbf{M}$ is fed into a BPNN network to realize the PCE value prediction. During the training process, mean square error (MSE) is used as the loss function to evaluate the performance of the regression model, and the model error is propagated backward to update the fragment-embedding matrix and the parameters in the Bi-LSTM and the attention layer. A stochastic gradient descent algorithm is selected as the loss function optimizer to minimize the MSE. A more comprehensive description of the Bi-LSTM network and attention mechanism is given in the following.

Bi-LSTM Network for Feature Integration on the Fragment Level. Bi-LSTM network is employed to obtain high-level molecular features from the molecular descriptor. It is an enhanced recurrent neural network (RNN) with the capability of processing lengthy sequences that takes the context information embedded in both the forward and backward sequence into consideration.[30,31] It has been widely used in the fields where the embedded information sequences span long intervals, such as unconstrained handwriting recognition,[32] machine translating,[33] and image captioning.[34] For a given molecule p, the corresponding descriptor is represented as $[\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_n]$ (here, n is fragment sequence dimension and x_t is the fragment vector). As shown in Figure 3b, each fragment vector for the present molecule is the input of a forward-directional and backward-directional LSTM cell, and the processed fragment information $\overrightarrow{\mathbf{h}_t}$ / $\overleftarrow{\mathbf{h}_t}$ is passed on to the next cell. For each LSTM cell, an adaptive gating mechanism is introduced to decide the degree to which the information of the previous fragment passed on by the previous cell is kept, and the features of the current fragment information input is memorized.[30]

A forward-directional LSTM cell is shown in Figure 3c. As shown, three gate units are introduced in the cell to control the information flow, a forget gate unit f_t, an input gate unit i_t, and an output gate unit o_t. All of these gates with the current information generated by the cell, g_t,control both the state of the cell, c_t, and that of the hidden unit, $\mathbf{h}_t$. The information-processing procedures of these gate units and cell states subject to eqs 2−6, during which a series of weight matrices $\mathbf{W}_{xi}$, $\mathbf{W}_{hi}$, $\mathbf{W}_{xf}$, $\mathbf{W}_{hf}$, $\mathbf{W}_{xc}$, $\mathbf{W}_{hc}$, $\mathbf{W}_{xo}$, and $\mathbf{W}_{ho}$ and bias parameters $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_c$, and $\mathbf{b}_o$ are applied to process the information input of the current cell $\mathbf{x}_t$, and the state information generated by the previous cell $\overrightarrow{\mathbf{h}_{t-1}}$. The processed information after the gate units is then used to update the current cell state c_t and its hidden state output $\overrightarrow{\mathbf{h}_t}$,[35] the formulations of which are given in eq 7.

$$i_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\overrightarrow{\mathbf{h}_{t-1}} + \mathbf{b}_i) \quad (2)$$

$$f_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\overrightarrow{\mathbf{h}_{t-1}} + \mathbf{b}_f) \quad (3)$$

$$g_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\overrightarrow{\mathbf{h}_{t-1}} + \mathbf{b}_c) \quad (4)$$

$$c_t = i_t g_t + f_t c_{t-1} \quad (5)$$

$$o_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\overrightarrow{\mathbf{h}_{t-1}} + \mathbf{b}_o) \quad (6)$$

$$\overrightarrow{\mathbf{h}_t} = o_t \tanh(c_t) \quad (7)$$

The Bi-LSTM network applied in this work contains n forward-directional LSTM cells and n backward-directional LSTM cells. After the information-processing procedure, a

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

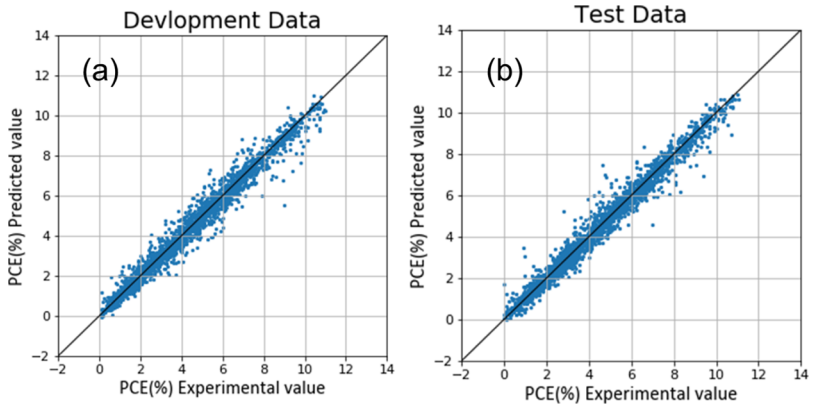


Figure 4. Scatter plots for the predicted experimental value with the QSPR model.

collection of hidden states is obtained for both forward and backward information extraction. The information extraction of the Bi-LSTM network is summarized in eqs 8 and 9.

$$\overrightarrow{\mathrm{LSTM}}([\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n]) = [\overrightarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2}, \ldots, \overrightarrow{\mathbf{h}_t}, \ldots, \overrightarrow{\mathbf{h}_n}] \tag{8}$$

$$\overleftarrow{\mathrm{LSTM}}([\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n]) = [\overleftarrow{\mathbf{h}_1}, \overleftarrow{\mathbf{h}_2}, \ldots, \overleftarrow{\mathbf{h}_t}, \ldots, \overleftarrow{\mathbf{h}_n}] \tag{9}$$

The element-wise sum is used to combine the outputs passed by the backward-directional cells $[\overleftarrow{\mathbf{h}_1}, \overleftarrow{\mathbf{h}_2}, \ldots, \overleftarrow{\mathbf{h}_n}]$ and the outputs passed by the forward-directional cells $[\overrightarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2}, \ldots, \overrightarrow{\mathbf{h}_n}]$, which is given in the following equation. The obtained information vector $\mathbf{h}_t$ is then used as an input for the subsequent attention layer.

$$\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t} \oplus \overleftarrow{\mathbf{h}_t}] \tag{10}$$

Attention Layer for Feature Extraction on the Molecule Level. Not all the component fragments contribute equally to the PCE performance of an organic compound. Therefore, the attention neural network is applied to track the fragments that are important to the PCE performance. Features extracted on the fragment level are merged into a molecule-level feature vector, by multiplying a normalized importance weight vector. The processed information of fragments, $\mathbf{h}_t$, is fed to a one-layer MLP (multilayer perceptron) to obtain $\mathbf{u}_t$, where the weight vectors, $\mathbf{W}_s$, and bias parameters, $\mathbf{b}_s$, are introduced, after which the obtained result was further used to generate a normalized importance weight α_t through a softmax function. After this, a weighted sum of the information vector is obtained as the processed molecular information, $\mathbf{M}$. The calculation formulations are given in the following.[36]

$$\mathbf{u}_t = \tanh(\mathbf{W}_s\mathbf{h}_t + \mathbf{b}_s) \tag{11}$$

$$\alpha_t = \frac{\exp(\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_s)}{\sum_i \exp(\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_s)} \tag{12}$$

$$\mathbf{M} = \sum_t \alpha_t \mathbf{h}_t \tag{13}$$

where $\mathbf{u}_s$ is a fragment context vector randomly initialized and jointly learned during the network training process, so are the other involved vectors, $\mathbf{W}_s$ and $\mathbf{b}_s$.

BP Neural Network for Property Prediction Based on the Processed Molecular Information. The three-layer BPNN, which contains one input layer, one hidden layer, and one output layer, is utilized to correlate the processed molecular information $\mathbf{M}$ and the PCE value of each molecule. The structure of the BPNN is also shown in Figure 3b. The input layer receives the information, $\mathbf{M}$, and the output layer gives the estimated PCE values.

Model Evaluation. The predictive performance, competitiveness and transferability of the developed QSPR model are assessed. A test data set is used to evaluate the prediction accuracy. The external competitiveness of the developed model is assessed by comparing it to other eight predictive models.

All the above steps for model training and evaluating are accomplished by a series of programs written in Python language and executed on both Windows and Linux platforms. An open-source deep-learning framework Pytorch[37] is used to implement the neural network, and the procedure is accelerated on two GTX-1080Ti GPUs.

RESULTS AND DISCUSSION

Fragment Information of the Collected Molecules. The molecules in the collected CEPDB data set are made up of 56 types of fragments, which are composed of seven element species, C, H, O, N, S, Si, and Se, and three chemical bonds, single bond, double bond, and aromatic bond. More details about the fragments are available in the Supporting Information.

Experimental Settings. The processed CEPDB data set containing 23,878 both SMILES and valid PCE values of organic compounds is randomly divided into three sets, a training set, a development set, and a test set according to their corresponding proportions (80, 10 and 10%). The development data set (2388 compounds) is used for the validation of model hyperparameters optimization during the model training procedure, while the test set (2388 compounds) is used for final model evaluation.

Grid search is applied to tune the model hyperparameters, including the selection of loss function optimizer, the learning rate, and the number of hidden layers units for the hidden layer of the BPNN and that for the Bi-LSTM layer. Adam[38] is chosen as the loss function optimizer with a learning rate of 0.001. After the model training and developing process, an

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

Table 1. Prediction Accuracy of the Testing Set

index	measure	MSE	R^2	index	measure	MSE	R^2
1	mordred + RF	0.56	0.90	4	mordred + ANN	0.55	0.90
2	sum-g-FSI + RF	0.26	0.95	5	sum-g-FSI + ANN	0.27	0.95
3	ECFP + RF	0.26	0.95	6	ECFP + ANN	0.26	0.95
7	g-FSI + transformer	0.85	0.85				
8	g-FSI + RNN	0.28	0.95				
9	g-FSI + Bi-LSTM + attention	0.12	0.98				

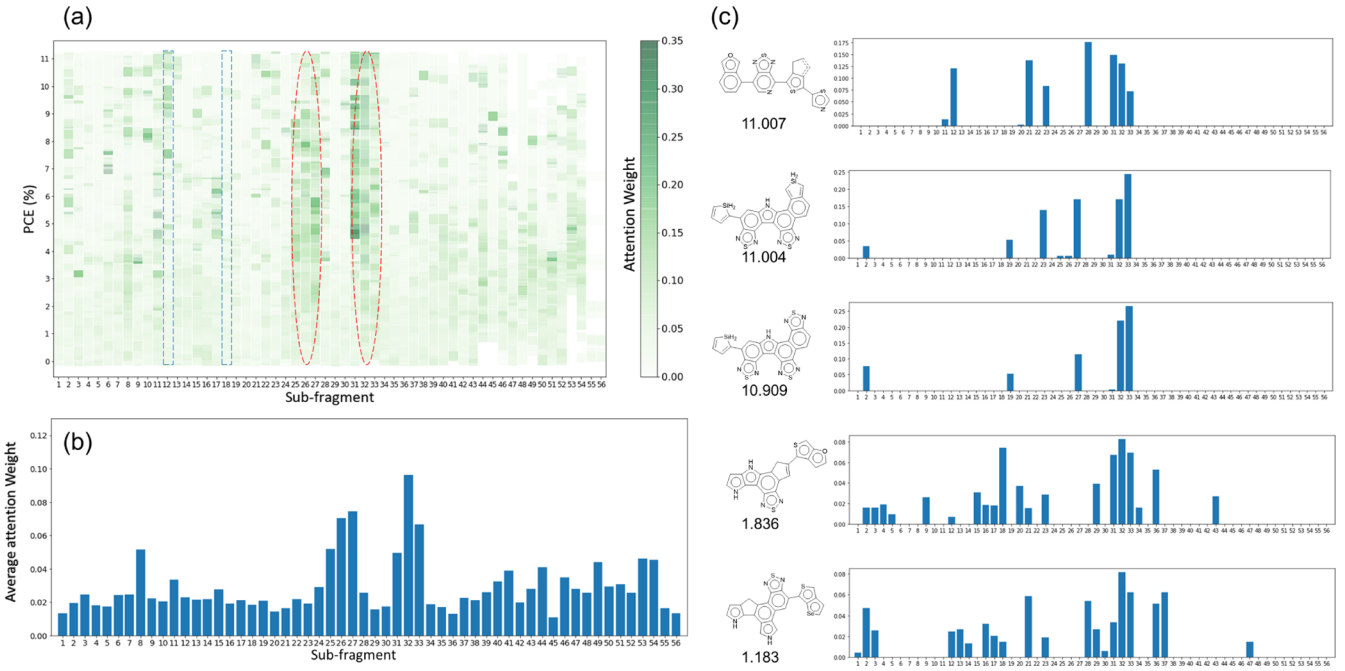


Figure 5. Visualization of the attention mechanism: (a) attention alignment map for the 56 organic solar cell fragments; (b) average attention alignment of the various fragments; and (c) attention alignment map for five concrete cases.

optimized QSPR model is obtained. Optimized hidden dimensions are 128 for Bi-LSTM layers and 32 for the hidden layer of the BPNN. Then, the test set is used to evaluate the obtained QSPR model. Figure 4 gives the comparison between the predictive PCE value by the obtained model and calculated theoretical value. For both the development and the test data set, it can be observed that, the predictive results are closely distributed along the diagonal line. The calculated Pearson's correlation coefficient for the obtained QSPR model is 0.98, and the prediction MSE for the development set and the test set are ±0.13 and ±0.12, respectively. Therefore, it can be concluded that the obtained QSPR model achieved high accuracy.

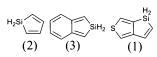
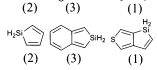
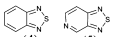
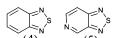
Competitiveness of the Model. In order to justify the rationality and reliability of the proposed approach, several other molecular descriptors and machine learning methods are also applied to develop the QSPR model using the same data set. The other selected molecular descriptors include the extended-connectivity fingerprints (ECFP)[39] and the mordred descriptors,[40] and the selected machine-learning algorithms are the ordinary RNN, the transformer,[36] ANN, and RF. Because the proposed molecular descriptor (g-FSI) is represented as a matrix, it cannot be taken as input for ANN and RF directly. For the convenience of comparison, it is converted into a vector by summing up the fragment vectors, and the resulted vector (sum-g-FSI) is used as the input for ANN and RF. For ECFP and mordred, they are both vectors of fixed length; thus, the generated descriptors are directly used as input for ANN and RF. In total, nine QSPR models are obtained. Table 1 presents the performance comparison of the QSPR models.

It can be seen from the comparison among the three molecular descriptors (sum-g-FSI, ECFP, and mordred) that a higher correlation coefficient, 0.95, and lower MSE, 0.26 and 0.27, are reached for sum-g-FSI and ECFP, suggesting that the proposed descriptor sum-g-FSI and ECFP can better suit the need of the current study. According to the result, no winner can be determined between sum-g-FSI and ECFP. This is because, essentially, they are the same type of molecular descriptor based on molecular fragment information, while for the two machine-learning methods, RF and ANN, equivalent modeling capacity is exhibited in this case. For sum-g-FSI and ECFP, the calculated correlation coefficient, MSE, and mean absolute error (MAE) are nearly equal for models developed from RF and ANN. Although slight differences are observed, it is most probably a result of the uncertainty in RF modeling.

For sum-g-FSI and g-FSI, the former is the compressed representation of the latter. It ignores the fragment sequence information of a molecule. Still, interestingly, the combination of g-FSI and traditional RNN did not outperform the sum-g-FSI and RF/ANN, implying that the traditional RNN failed to make effective utilization of the fragment sequence information embedded in the proposed descriptor; this is owing to the fact that traditional RNN cannot effectively deal with the problems of vanishing or exploding gradient.[30] The result indicates that

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

Table 2. Relationship between Fragments and Building Blocks

Index	Fragment	Building blocks containing the sub-fragment	Amount of molecules containing those building blocks	Average PCE value	Deviation from standard value*
8		 (1)	387	6.04	+1.92
26		 (2) (3) (1)	4106	5.77	+1.65
27		 (2) (3) (1)	4106	5.77	+1.65
32		 (4) (5)	5880	6.59	+2.47
33		 (4) (5)	5880	6.59	+2.47

Fragment group	Amount of molecules containing the fragment group	Average PCE value	Deviation from standard value*
26, 27, 32, 33	784	7.78	+3.66
8, 26, 27, 32, 33	67	8.06	+3.94

Bi-LSTM can effectively extract the positional information of molecular fragments embedded in the descriptor and support better QSPR modeling. The model correlation coefficient has increased to 0.98 and MSE reduced to 0.12, by applying the Bi-LSTM.

As for transformer, it is chosen in this study because it can also take matrices as inputs and has exhibited widely accepted performance in natural language processing in recent years, but it did not deliver satisfactory performance in this case as expected.

Attention Mechanism for Decisive Fragments Analysis. For material design, result interpretability is no less critical than the predictive accuracy of a machine-learning model.[41,42] Similar to the experts' empirical intuitions and/or experience, information gained during the model "learning" process can be instructive for better OPVs design. For OPV design, one crucial part is to generate novel molecules, which usually requires researchers to offer predefined fragments and chemical design rules based on their experiences and knowledge.[43] This section analyzes the important fragments that are beneficial for the potential PCE performance of organic compounds, based on the attention weights granted to each molecular fragment by attention mechanism via the "learning" process. Therefore, the crucial predefined fragments could be searched automatically without additional researcher intuition.

For a molecular fragment, the more important it is, the higher attention weight it will gain in the model training. Figure 5a illustrates the heat map of the attention weight that the 56 fragments gained over the organic compounds with different PCE in the training data set. The darker the color, the higher the attention weight. It can be observed that there are two fragment groups that have received noticeable attention in most of the organic compounds, namely fragment 25 ∼27 and 31 ∼33. As shown in Figure 5b, the average attention weights for these fragments are also higher than that of the rest.

Five compounds containing fragments 32 and 33 are extracted from the test set (shown in Figure 5c). Interesting observations can be derived from the comparison of the fragment constitution of these five compounds. For the first three molecules with PCE values higher than 10.5%, their performances are mainly influenced by either the fragments with high average attention weight or the fragments of relatively lower weight but mostly observed in high PCE compounds. For instance, the PCE performance of the first compound is mostly affected by fragment 12, 21, 23, 28, and 31 ∼33. Apart from fragment 31 ∼33, the other four fragments all showed high impact in compounds with PCE higher than 5%. More specifically, if one check the heat map distribution for fragment 12, one can find that the color coded on the upper side above PCE = 6% is notably darker. While for the remaining two compounds with PCE values lower than 2%, it is easy to note that another type of fragment also played an important role. Take the fourth compound as an example, one of the important contributor fragment is 18, of which the overall attention weight distribution shown in the heat map is mostly light colored. Besides, the two with lower PCEs are in a more average distribution of attention weights than the other three. We infer that fragments such as 18 may deteriorate the potential PCE performance of compounds.

With these findings, let us take a further step to find out the fundamental function fragments From Figure 5b, the top five fragments with the highest average attention weight can be singled out, which are fragments 8, 26, 27, 32, and 33. More detailed information of these five fragments is listed in Table 2. Intriguingly, we noticed that the first three fragments (8, 26, and 27) are all key constructive fragments for silole and its derivatives, whose characteristic electronic structures have been proved to be attractive in OPV applications.[44,45] The remaining two fragments (32 and 33) are commonly featured in their sulfur and nitrogen composition. They are the key building fragments of benzothiadiazole, which is another proven candidate unit for high-efficiency OPV materials because of the large conjugate plane and strong electron-withdrawing properties.[46]

Moreover, the building blocks containing any of the five top fragments are identified from the building block pool. Here, the building block pool refers to the collection of building blocks that made up the organic compounds in CEPDB,[13] for which a more detailed description is given in the Supporting information. As given in Table 2, five such building blocks are identified, which are 1-*H*-siolo[2,3,-*c*]thiophene (1), silacyclopenta-2,3-dine (2), 2*H*-2-silaindene (3), benzothiadiazole (4), and [1,2,5]-thiadiazolo[3,4,-*C*]pyridine (5). Following the identification of the five key building blocks, all the organic compounds that contain such building blocks are extracted and their average PCE values calculated. The average PCE performance of these organic compounds are all higher than

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

the average level. It is worth noting that, interestingly, these findings agree very well with that of Hachmann and co-workers' study.[16] In their study, four out of these five identified building blocks were found to be strongly overexpressed in the top PCE set. This proved that the proposed molecular description language is effective, and the established deep-learning method is capable of "learning" essential chemical information automatically.

Furthermore, we studied the molecules that contain more than three of the top five fragments. It is found that the average PCE value calculated for these organic compounds are much higher than that of which containing lesser type of fragments (as illustrated in Table 2). More interestingly, the compounds that contain all the top five fragments have the highest average PCE value (8.06), suggesting that molecules with more of the key fragments are more likely to be better OPV candidates.

Transferability of the Proposed Method to External Data Sets. Another two data sets collected from NFADB[15] and HOPV15[14] are used to assess the transferability of the proposed method. Seven element species (*i.e.* C, H, O, N, S, Si, and F), and 196 types of fragments are involved in the formation of the organic compounds in the data set collected form the former database. For that of the latter, 133 types of fragments and one more element species, Se, is involved.

The collected data sets (3000 from NFADB and 373 from HOPV15, respectively) are proportionally divided into two parts, a training set (80%) and a test set (20%), which are then used to train and test the proposed methodology, in order to study its transferability. For the purpose of methodology comparison, another two molecular descriptor (mordred and ECFP) and one machine learning method (ANN) are also applied to these two data sets. Comparison of the MAE between the proposed model and the other two models (mordred + ANN and ECFP + ANN, respectively) is given in Figure 6. It can be seen that, unsurprisingly, for all the three

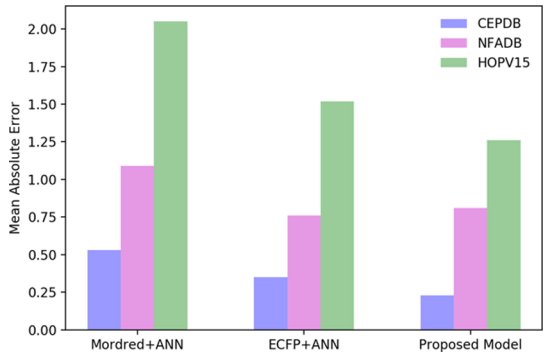


Figure 6. MAE for the CEPDB, NFDB, and HOPV15 data.

methods, the model accuracy declines along with the decrease in the number of training data sets. For all the three data sets, in most of the cases, the proposed method outperforms the other two methods. When trained by the data sets from NFADB, the MAE of the proposed method is slightly higher than that of "ECFP + ANN". This is probably because of the fact that the calibration of data from NFADB has employed ECFP. Therefore, it can be concluded from the comparison results that the proposed molecular descriptor is an effective representation for organic compound description; and the established deep-learning network is capable of delivering highly accurate property prediction from the proposed molecular descriptor.

CONCLUSIONS

Photovoltaic technology is considered as one of the most promising ways to resolve the energy shortage and environmental crisis in this century. The discovery of chemical compounds with high potential photoelectric conversion efficiency has become one of the pivotal tasks to push this technology forward. This paper proposes an interpretable methodology to develop an accurate and reliable QSPR model for OPV materials. Inspired by the similarity between organic chemistry and natural language, a language-like molecular descriptor is proposed to describe organic compounds and a deep-learning model is established to realize high-accuracy PCE value prediction. The fragments consisting of atoms and bonds are embedded to numeric vectors, and the related vectors are aggregated into a matrix based on the sequence information of the fragments. It is believed that similar to a sentence, the position information of the fragments (words) is important to the property prediction (meaning comprehension) of the compound (sentence). Bi-LSTM is employed to process the molecular descriptor, such that the embedded fragment sequence information can be fully "comprehended". Subsequently, the processed information is fed to a BPNN to realize PCE value prediction. During the procedure, the attention mechanism is applied to help recognize the significance of fragments and improve predictive accuracy. Model evaluation results indicate that higher predictive accuracy can be reached and better transferability to other data sets is observed for the proposed model, in contrast to several other molecular descriptors and machine learning algorithms. Result comparison also proves that sequence information of the fragments is important to the potential PCE performance prediction of organic compounds, and such information can be captured by the improved recursive neural network. Moreover, the developed methodology can, to some extent, reveal the contributions of the fragments to the PCE performance of a molecule, which may be helpful for the reverse design of OPVs.

Both the descriptor generation and property-mapping procedures are accomplished automatically, which avoids human intervention during the whole process. That is to say, the deep-learning approach is capable enough of extracting and learning important knowledge from SMILES, and thus, no in-depth domain knowledge is required from the modeler. Furthermore, several decisive fragments and the corresponding building blocks are identified during the "learning" process of the proposed method, indicating that the proposed method is capable of providing instructive information for the reverse design of OPVs. Although this study focuses on the PCE value prediction of OPVs, the proposed approach can be further extended to the prediction of other important properties for organic materials because of the similarity between organic chemistry and English. As for the property estimation of inorganic molecules and polymers, this model might not give a satisfying predictive result without additional modification.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.0c03880.

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000

Fragments of organic compounds in CEPDB, NFADB, and HOPV15 and the procedure of the BFS algorithm (PDF)

AUTHOR INFORMATION

Corresponding Author

Li Zhou − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China;* orcid.org/0000-0003-3539-7573; Email: chezli@scu.edu.cn

Authors

Jinkui Wu − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China;* orcid.org/0000-0002-4839-338X

Shihui Wang − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China*

Xu Ji − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China*

Yiyang Dai − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China*

Yagu Dang − *School of Chemical Engineering, Sichuan University, Chengdu 610065, China;* orcid.org/0000-0003-1715-3153

Markus Kraft − *Cambridge Center for Advanced Research and Education in Singapore Ltd., 138602, Singapore; Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, U.K.; School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459 Singapore;* orcid.org/0000-0002-4293-8924

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.0c03880

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Financial support from the Fundamental Research Funds for the Central Universities (YJ201838) and the National Natural Science Foundation of China (21776183, 21706220) is gratefully acknowledged.

REFERENCES

(1) Srinivasan, M. V.; Ito, M.; Kumar, P.; Abhirami, K.; Tsuda, N.; Yamada, J.; Shin, P.-K.; Ochiai, S. Performance Evaluation of an Organic Thin-Film Solar Cell of PTB7:PC 71 BM with an Alcohol-Soluble Polyelectrolyte Interlayer Prepared Using the Spray-Coating Method. *Ind. Eng. Chem. Res.* **2015**, *54*, 181−187.

(2) Roy, A.; Ghosh, A.; Bhandari, S.; Sundaram, S.; Mallick, T. K. Realization of Poly(methyl methacrylate)-Encapsulated Solution-Processed Carbon-Based Solar Cells: An Emerging Candidate for Buildings' Comfort. *Ind. Eng. Chem. Res.* **2020**, *59*, 11063−11071.

(3) Wang, Z.; Su, Y.; Shen, W.; Jin, S.; Clark, J. H.; Ren, J.; Zhang, X. Predictive deep learning models for environmental properties: the direct calculation of octanol−water partition coefficients from molecular graphs. *Green Chem.* **2019**, *21*, 4555−4565.

(4) Liang, X.; Zhang, X.; Zhang, L.; Liu, L.; Du, J.; Zhu, X.; Ng, K. M. Computer-Aided Polymer Design: Integrating Group Contribution and Molecular Dynamics. *Ind. Eng. Chem. Res.* **2019**, *58*, 15542−15552.

(5) Hu, H.; Yuan, Z. Hard-threshold neural network-based prediction of organic synthetic outcomes. *BMC Chem. Eng.* **2020**, *2*, 7.

(6) Chai, S.; Liu, Q.; Liang, X.; Guo, Y.; Zhang, S.; Xu, C.; Du, J.; Yuan, Z.; Zhang, L.; Gani, R. A grand product design model for crystallization solvent design. *Comput. Chem. Eng.* **2020**, *135*, 106764.

(7) Scharber, M. C.; Mühlbacher, D.; Koppe, M.; Denk, P.; Waldauf, C.; Heeger, A. J.; Brabec, C. J. Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Adv. Mater.* **2006**, *18*, 789−794.

(8) Green, M. A.; Dunlop, E. D.; Hohl-Ebinger, J.; Yoshita, M.; Kopidakis, N.; Ho-Baillie, A. W. Y. Solar cell efficiency tables (Version 55). *Prog. Photovoltaics Res. Appl.* **2020**, *28*, 3−15.

(9) Gómez-Bombarelli, R.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120−1127.

(10) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **2017**, *3*, 53.

(11) Wu, S.; Kondo, Y.; aki Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **2019**, *5*, 66.

(12) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.

(13) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.

(14) 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*, 160086.

(15) 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*, 857−870.

(16) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry − the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698−704.

(17) Padula, D.; Simpson, J. D.; Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **2019**, *6*, 343−349.

(18) 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*, 1801032.

(19) 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*, 17480−17488.

(20) 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 Simul.* **2019**, *2*, 1800116.

(21) Paul, A.; Jha, D.; Al-Bahrani, R.; Liao, W.-k.; Choudhary, A.; Agrawal, A. Transfer Learning Using Ensemble Neural Networks for Organic Solar Cell Screening. *Proceedings of the International Joint Conference on Neural Networks 2019*, 2019-July, 1−8.

(22) Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A. Organic Chemistry as a Language and the Implications of Chemical Linguistics for Structural and Retrosynthetic Analyses. *Angew. Chem., Int. Ed.* **2014**, *53*, 8108−8112.

(23) David, W. SMILES: A chemical language and information system. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(24) Aksoy, S.; Haralick, R. M. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recogn. Lett.* **2001**, *22*, 563−582.

(25) RDKit: Open-source cheminformatics. 2020, http://www.rdkit.org; [Online; accessed April 2, 2020].
(26) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference*: Pasadena, CA USA, 2008; pp 11−15.
(27) Costa, F.; De Grave, K. Fast neighborhood subgraph pairwise distance kernel. *ICML 2010—Proceedings, 27th International Conference on Machine Learning*, 2010; pp 255−262.
(28) Meghanathan, N. *Routing Protocols and Graph Theory Algorithms for Mobile Ad Hoc Networks*; IGI Global, 2017; pp 971−1411.
(29) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. **2014**, arXiv preprint arXiv:1409.0473.
(30) Graves, A. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.
(31) Schuster, M.; Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673−2681.
(32) Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 855−868.
(33) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *4*, 3104−3112.
(34) Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image Captioning with Deep Bidirectional LSTMs. *Proceedings of the 2016 ACM on Multimedia Conference—MM '16*: New York, New York, USA, 2016; pp 988−997.
(35) Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*: Stroudsburg, PA, USA, 2016; Vol. 2: Short Papers, pp 207−212.
(36) Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc., 2017; pp 5998−6008.
(37) Paszke, A. et al. *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp 8024−8035.
(38) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**, arXiv: preprint arXiv:1412.6980.
(39) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(40) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.
(41) Bang, S.; Xie, P.; Lee, H.; Wu, W.; Xing, E. Explaining a black-box using Deep Variational Information Bottleneck Approach. *35th International Conference on Machine Learning, ICML 2018*, 2019; Vol. 2, pp 1386−1418.
(42) Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **2018**, *61*, 36−43.
(43) Kim, K.; et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.* **2018**, *4*, 67.
(44) Yamaguchi, S.; Tamao, K. A Key Role of Orbital Interaction in the Main Group Element-containing π-Electron Systems. *Chem. Lett.* **2005**, *34*, 2−7.
(45) Zhang, Z.; Zhu, X. Bis-Silicon-Bridged Stilbene: A Core for Small-Molecule Electron Acceptor for High-Performance Organic Solar Cells. *Chem. Mater.* **2018**, *30*, 587−591.
(46) Li, G.; Kang, C.; Gong, X.; Zhang, J.; Li, W.; Li, C.; Dong, H.; Hu, W.; Bo, Z. 5,6-Difluorobenzothiadiazole and silafluorene based conjugated polymers for organic photovoltaic cells. *J. Mater. Chem. C* **2014**, *2*, 5116−5123.

https://dx.doi.org/10.1021/acs.iecr.0c03880
Ind. Eng. Chem. Res. 2020, 59, 18991−19000