JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

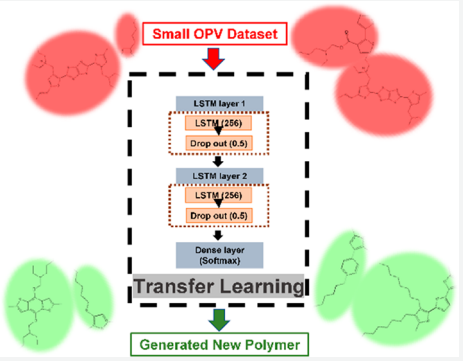
Transfer Learned Designer Polymers For Organic Solar Cells

Joydeep Munshi, Wei Chen, TeYu Chien, and Ganesh Balasubramanian*

Cite This: J. Chem. Inf. Model. 2021, 61, 134−142 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Organic photovoltaic (OPV) materials have been examined extensively over the past two decades for solar cell applications because of the potential for device flexibility, low-temperature solution processability, and negligible environmental impact. However, discovery of new candidate OPV materials, especially polymer-based electron donors, that demonstrate notable power conversion efficiencies (PCEs), is nontrivial and time-intensive exercise given the extensive set of possible chemistries. Recent progress in machine learning accelerated materials discovery has facilitated to address this challenge, with molecular line representations, such as Simplified Molecular-Input Line-Entry Systems (SMILES), gaining popularity as molecular fingerprints describing the donor chemical structures. Here, we employ a transfer learning based recurrent neural (LSTM) model, which harnesses the SMILES molecular fingerprints as an input to generate novel designer chemistries for OPV devices. The generative model, perfected on a small focused OPV data set, predicts new polymer repeat units with potentially high PCE. Calculations of the similarity coefficient between the known and the generated polymers corroborate the accuracy of the model predictability as a function of the underlying chemical specificity. The data-enabled framework is sufficiently generic for use in accelerated machine learned materials discovery for various chemistries and applications, mining the hitherto available experimental and computational data.



INTRODUCTION

Organic solar cells (OSC) are lightweight, flexible, and inexpensive than the inorganic silicon photovoltaic devices.[1−3] Nevertheless, the limited power conversion efficiency (PCE) of bulk heterojunctions (BHJ) OSCs compared to silicon-based solar cells has posed a mammoth challenge toward their large-scale commercialization.[4−10] Although recent efforts have enabled a maximum cell efficiency of ∼18%,[11−15] commercial OSC devices seldom achieve a PCE beyond 10% and the upscaling of OSCs from lab prototypes to fabricated devices is substantially constrained by the inherent instability of the employed materials.[16]

Semiconducting conjugated polymers are the predominantly preferred donor materials for OSC devices due to their exceptional optoelectronic properties emerging from the mobile π-electrons.[17] Together with fullerene-based acceptor materials the conjugated polymers have benchmarked the potential ability of OSCs.[1,2,18,19] Conventionally, design and synthesis of conjugated polymer materials with desired optoelectronic properties have been guided by recursive experimental synthesis, characterization, and optimization of the device performance. To complement such experimental efforts that consider an experiential choice of materials, computer simulations based on first-principles, classical molecular dynamics (MD), coarse-grained MD, and other computational frameworks have been leveraged.[5−7,10,20−28] However, these modeling techniques are restricted to length and time scales that are much displaced from the physical experiments, as well as to selection of specific materials based on the availability of molecular interaction functions and parameters. The emergence of machine learning (ML) methods for material science[29] holds promise to potentially overcome the above computational challenges and accelerate materials design.

ML has previously been employed for data-enabled materials discovery for organic electronic and photovoltaics.[30−35] ML techniques are typically classified into (i) supervised and (ii) unsupervised learning approaches. For instance, supervised ML has been applied for the use of random forest (RF) classifier to predict candidate polymers with enhanced PCE based on the molecular information embedded in their unique fingerprints.[36] In other reports, regression techniques, such as the ensemble-based regression,[37] RF and extremely randomized trees[38] were implemented to predict molecular orbital energies from the corresponding molecular fingerprints. Jørgensen et al.[39] demonstrated a deep generative scheme to predict molecular properties via context free grammar variational auto encoders (VAE). Recent advances in natural language processing (NLP)[40−42] and machine translation[43] demon-

Received: October 5, 2020
Published: January 7, 2021

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

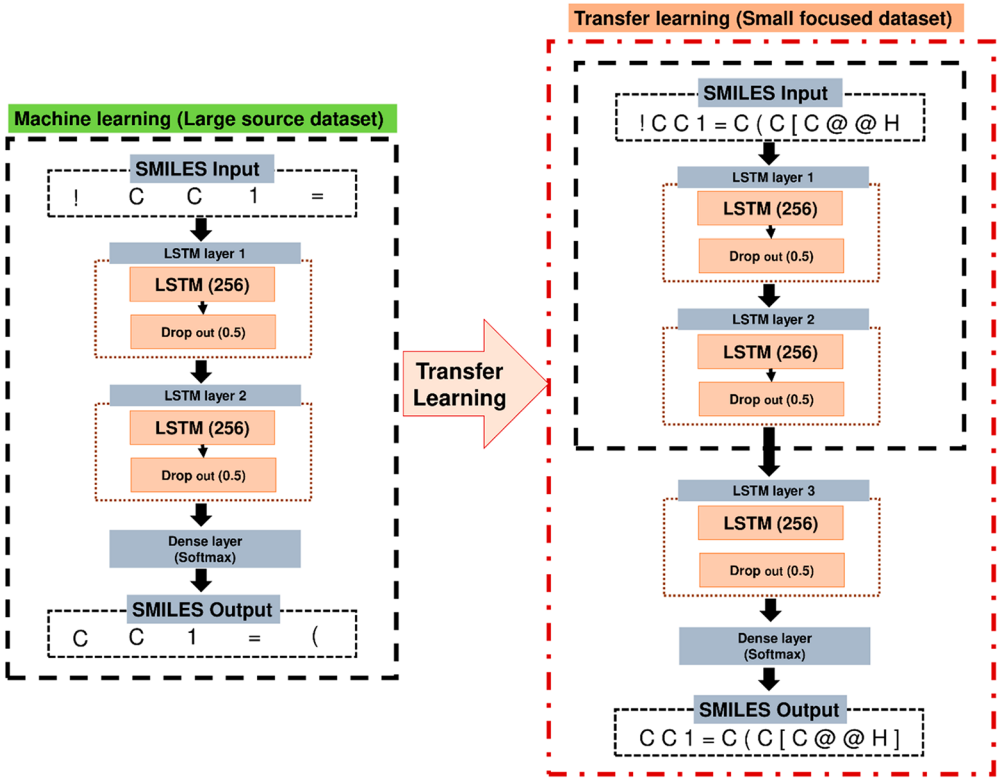


Figure 1. Flowchart depicting the LSTM model and the transfer learning predictive scheme. The red dashed line (right) represents the pretrained LSTM combined with the new LSTM layer augmented to fine-tune the model. Initially the RNN model with 2 layers of LSTMs (left) is trained on a large data set of organic molecules to learn the chemical grammar from SMILES notations. Once the model converges on the large data set, transfer learning approach is implemented by augmenting a new LSTM layer 3 over the pretrained network. The final LSTM model is trained on a small focused data set to generate new polymer repeat units for OPV applications.

strated the remarkable success of recurrent neural network (RNN) in a gamut of applications from socioeconomic challenges to materials design problems.[44−47]

Long short-term memory (LSTM) cells[48] are the most attractive candidates to successfully achieve sequence to sequence predictions and time series analysis and are extensively employed to predict properties of proteins[49] and solubility of druglike compounds.[50] Predictive models using connectivity rules for molecular structures, such as Simplified Molecular-input Line-entry System (SMILES), International Chemical Identified (InChI), and Molecular Access System (MACCS), are progressively becoming popular in chemical and bioinformatics. Prior literature demonstrated the ability of RNNs to generate canonical SMILES strings spanning a wide spectrum of chemical space.[51,52] Transfer learning, on the other hand, has enabled the flow of information from an already learned task to a relatively unexplored activity[30,40,53] and hence can be successfully employed to predict results from a limited training data set, such as those that exist for OSC materials.

Inspired by the transfer learned de novo drug design,[51] here, we construct a transfer learning predictive scheme based on LSTM deep neural network (DNN) to generate SMILES of polymers as candidate materials for OSC devices. Initially, the LSTM model is established to generate new SMILES from a large data set of known small organic molecules (∼1 million). Subsequently, the pretrained network is further perfected on a relatively smaller size data set (∼1400 conjugated polymers) to generate new polymer repeat units that are structurally similar to typical OPV donor materials. The generated molecules are further validated by predicting electrical properties, such as PCE, fill factor (FF), molecular orbital, and bandgap energies of the associated donor molecules based on the extracted molecular descriptors. The choice of descriptors is visualized using a principal component analysis (PCA) and RF regression is implemented to design a supervised learning model that predicts properties of new molecules from the unique molecular descriptors embedded in the generated SMILES.

DATA-ENABLED AND COMPUTATIONAL TECHNIQUES

Data Sets. The LSTM model is trained with ∼1 million SMILES strings of small organic molecules extracted from GDB17 chemical database,[54] which contains small organic druglike molecules of up to 17 atoms consisting of C, N, O, S, and halogens (F, Cl, Br, etc.). Initially the entire data set is prechecked for any duplicate entries followed by random shuffling of all the molecules. The SMILES strings extracted from the data set are used to train an LSTM model to learn the grammar of the chemical language. Finally, a smaller data set consisting of ∼1400 monomer repeat units of OPV donor materials is gathered (Supporting Information) from the Harvard Organic Photovoltaic (HOPV15) data set[55] and experiments[36] to fine-tune the pretrained LSTM model to generate candidate polymers. While the molecular orbital energies and the bandgap energies are obtained using first-principle calculations,[55] the electrical properties such as the

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

power conversion efficiency (PCE) and fill factor (FF) are gathered from several experimental efforts.

LSTM Model. RNN processes a series of information embedded in a string, such as $\mathbf{S} = S_1S_2S_3...S_n$, by operating on one input, S_i, at a time. During each training cycle, the sequential inputs through a series of gates (forget gate, input gate, and output gate in LSTM) are transported together with the hidden state vectors (h_i) from previous RNN units, and the output vectors ($\hat{Y}_i$) are returned when the corresponding flag is turned on. The hidden state, h_i, is the core component of any RNN model as it passes information related to what RNN has seen previously during the training. Here, we implement a class of RNN, that is, LSTMs, as they are able to supervise the flow of information within the RNN network and control which of the hidden states should be passed through the successive cells. In addition, presence of forget gates enable LSTMs to retain relevant information while assimilating the correlations and underlying context from long sequence of data. This setup enables LSTMs to predict the context of very long sequence of words or sentences without any problems during the back-propagation process. Figure 1 presents the RNN architecture implemented in this work.

The initial training on the large database is accomplished through two LSTM layers followed by dropout regularization to avoid any risk of overfitting. The size of the LSTM hidden layer (i.e., LSTM units) and the rate of dropout are considered as the hyperparameters in our model. We employ Sequential Model-Based Global Optimization (SMBO) algorithm using Tree-Structured Parzen Estimator (TPE)[56] implemented through Hyperopt package[57] to ensure the best model performance (Figure 2A). The TPE-based algorithm executed

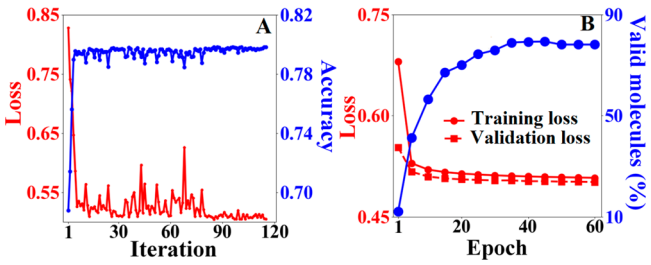


Figure 2. Hypermeter optimization and performance analysis of LSTM model. (A) Hyperparameter optimization of the LSTM model using SMBO algorithm. The overall loss (categorical cross-entropy) for each set of hyperparameter is obtained during each iteration and a global minimum is attained after ∼100 iterations. The 2 layers of LSTM with 256 units and dropout rate of 0.5 are obtained as the global solution. (B) Training, validation loss, and percentage of valid molecules as a function of epochs for C_D = 2.5. The training and validation losses converge after 25 epochs. However, the percentage of valid molecules sampled during the training is converged after ∼50 epochs of production run.

with the validation loss as an objective function attains the global minimum with 256 units for the LSTM units and a dropout regularization rate of 0.5. After the first round of training on the large data set, we fine-tune the LSTM framework using the transfer learning approach to further train the model on the smaller OPV data set. To realize the transfer learning, we augment the pretrained LSTM architecture with an extra LSTM layer consisting of 256 units and a dropout rate of 0.5 while preserving the pretrained weights for the preceding layers, as shown in Figure 1. For both the LSTM models, we consider one-hot encoded representation of the SMILES strings as input **S** and target vector Y_i. The RNN model is thus trained based on maximum likelihood estimation while the categorical cross-entropy is considered as the loss function. The output vector $\hat{Y}_i$ from the LSTM model is a probability distribution of all the possible tokens with an aim to attain the maximum probability assigned to the correct token.

SMILES Encoding and Sampling. To train the LSTM model, we initially pad every string to the longest (N) SMILES in the data set. Each padded input string is then prefixed with a token '!' and suffixed with a token 'E' to discern the beginning and ending of an input string respectively. Subsequently, the padded and appended input strings are encoded to one-hot vectors of size $N \times M$, where M denotes the size of the vocabulary of all the tokens. Given an input token the RNN model predicts the next token in the sequence, which is then compared to the target one-hot encoded vector (Y_i) to obtain the average loss. Once the training is completed for the trained RNN model to generate new and unique SMILES, a sampling method is implemented inspired by the sampling of druglike molecules by Gupta et al.[51] To start the generation of new molecules, we feed the RNN model with the start token '!' and sample the next possible token from its probability distribution. The process of predicting next token is continued until either a desired length of string is achieved or the end token 'E' is predicted. Each time the predicted character is concatenated to the preceding string and fed back to the input. Supplementary to the softmax function, an additional diversity coefficient C_D is introduced to modify the extent of the exploration. While higher C_D results in great structural diversity, through an extended exploration a low value of C_D leads to identical yet reliable predictions. Table 1 lists the percentage of valid molecules predicted based on nearly 50 000 generated SMILES trained on the large data set.

Table 1. Percentage of Valid Molecules Generated from LSTM Model Trained on the GDB17 Database and Their Tanimoto Similarity Coefficient (T_C)[a]

	C_D = 0.5	C_D = 1.5	C_D = 2.5
valid molecules (%)	98.90	94.60	78.10
similarity (T_C)	0.5336 ± 0.007	0.4644 ± 0.006	0.4153 ± 0.005

[a]Although higher value of C_D produces diverse and unique molecules, depicted by the reduction in T_C for high C_D, a low value of the coefficient results in reliable predictions.

Information Retrieval and Featurization. SMILES are one-dimensional representation of a 3-dimensional molecular structure using simple line-entry notations. SMILES strings are useful to represent molecules with their unique fingerprint embedded in a line notation. With a simple set of vocabulary and grammar rules, it is a true language that can be digested using a sequential deep learning algorithm. As SMILES string contains almost every information related to molecular structure and arrangements of atoms in a molecule, physiochemical properties can be extracted from a SMILES string. We use RDkit open source cheminformatics library[58] to extract descriptors such as molecular weight, average molecular density, partial charges, number of radical electrons, number of aromatic and aliphatic rings etc. We extract ∼20 descriptors and perform principal component analysis (PCA) on the training, test, and generated molecule data sets to compare the

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

extent of similarity between the predicted and actual set of molecules. To compute the electronic properties of a typical organic solar cell (OSC), such as power conversion efficiency (PCE), fill factor (FF), bandgap energy, and highest molecular orbital energy, we train a regression-based supervised learning model using random forest regressor, where all the extracted physiochemical descriptors are considered as features.

RESULTS AND DISCUSSION

We train the LSTM model on GDB17 data set with 1% of the data set (~10 000 data points) assigned for validation and testing. With the initially large data set, ~78% valid molecules with C_D = 2.5,~94% valid molecules with C_D = 1.5, and ~98% with C_D = 0.5 are generated as presented in Table 1. Figure 2B presents the validation and training losses during 50 epochs of production run. The number of valid molecules and the performance metrics converge suitably after 40 epochs. PCA analysis of the generated and actual SMILES are compared (Figure 3A) to validate that the generated molecules share the same chemical space of the actual molecules from the training data set. Furthermore, in Figure 3B, the distribution of molecular weights of the generated and actual SMILES are contrasted to corroborate the similarity in the overall size of the organic molecules. We find the overall distribution and median of the molecular weights to be similar for both the data sets. Although deep learning approaches, such as CNN-RNN,[30] and attention-based encoder-decoder[53] models have been previously employed to chemical text prediction, the LSTM model in this work is simpler to implement as a generative model with most of the one-dimensional chemical text data. We further quantify the accuracy of the predicted molecules by calculating the chemical similarity of known organic molecules against the generated ones.

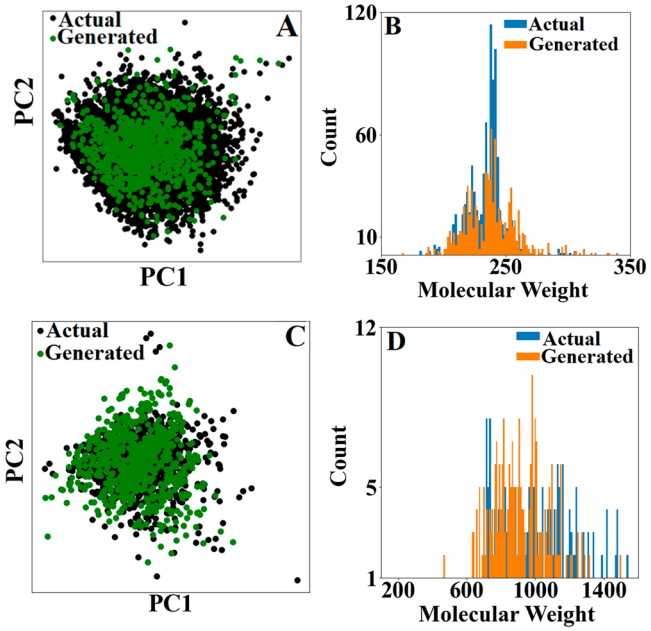


Figure 3. Information retrieval and feature interpretation from physiochemical descriptors. (A) A 2-dimensional principal component analysis (PCA) for known and generated molecules. Both actual and generated molecules reveal great similarity in the chemical space depicted by the PCA. (B) Distribution of the molecular weights (size of molecule) of the predicted structures. The median and overall distribution of the sampled molecules agree with the known data set. (C) PCA analysis of the sampled molecules from the transfer learned model. Similar to the previous LSTM model, the upgraded model successfully predicts new molecules spanning the same chemical space. (D) Molecular weight distribution of the predicted structures. Although the median is observed to shift to the left for the generated molecules, the overall distribution matches with the known data set.

Comparison of the molecular fingerprints in the form of SMILE strings are obtained from Tanimoto coefficient ($0 < T_C < 1$).[59] While the completely identical SMILES, according to Tanimoto similarity, attain $T_C = 1$, the dissimilarity between two molecules is represented by $T_C \sim 0$. The average Tanimoto similarity ~0.534 ± 0.006 for the generated sample data set with C_D = 0.5 confirms the predictability of the trained model as a function of the underlying chemical composition. As expected, increasing the diversity coefficient results in a reduction of the overall similarity between the molecular structures of the generated and the known organic species (Table 1). Figure 4 presents the chemical structure of a randomly nominated known molecule from GDB17 data set and its closest and furthest neighbors based on the T_C calculation. While the given molecular structure in Figure 4A resembles the chemical specificity of the generated molecule in Figure 4B because of the presence of the amino ($R\text{-}NH_2$) functional group, Figure 4C presents extreme divergence from the known molecule with respect to the functional group as well as the number of rings.

Once the training accuracy of the LSTM model for large data set is proven, we perfect the model using the transfer learning approach as shown in Figure 1. We continue the production run on the small focused data set of ~1400 conjugated polymer repeat units (monomers) for 25 epochs to train the new weights of LSTM layer 3 (Figure 1), while LSTM layers 1 and 2 are constrained to keep the pretrained weights. Once the overall loss function converges, further modification of all the three LSTM layers are performed for another 25 epochs with a very low learning rate ($\sim 10^{-5}$) to avoid any possible overfitting. A set of 1000 molecules are generated from the fine-tuned LSTM model with ~90% valid SMILES for C_D = 1.5. The reduction in the fraction of valid molecules is attributed to the long size of the SMILES strings of OPV materials. On the other hand, we observe <25% of the valid molecules if the transfer learned fine-tuning approach is not adopted. Next, we analyze the predictive accuracy of the underlying chemical specificity from the focused data set. Figure 3C evaluates the PCA results of the generated candidate polymer molecules against the OPV polymers from the known data set, revealing overlap between the chemical spaces from the physiochemical descriptors. Figure 3D further affirms the performance of the transfer learning from a comparative scrutiny of the distribution of molecular weights, similar to Figure 3B. In general, conjugated polymers used in OPV applications have similar order of magnitude ($\sim 10^3$ g/mol) for the molecular weight of the monomer units. Although the median and distribution of the total molecular size is observed to shift toward lower value than that of known molecules, the overall overlap of densely populated region between molecular weight of 600 and 1000 g/mol. reveal the ability of the fine-tuned model to predict the molecular size of polymer repeat units. Additionally, we compare quantitatively the chemical similarity of our updated model using the Tanimoto coefficient, as described above. The average $T_C \sim 0.4225 \pm 0.010$ corroborates the accuracy of the transfer learning approach presented here. We further verify the potential use

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

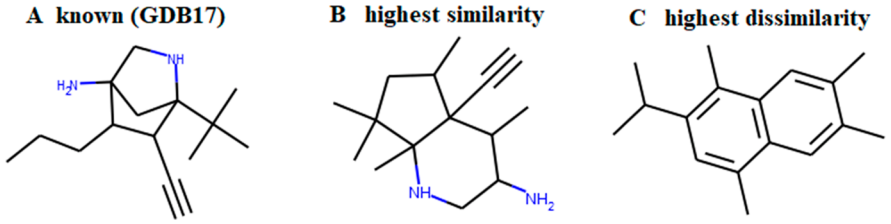


Figure 4. Comparison of generated molecules with known samples from GDB17 data set. (A) The chemical structure of a known organic molecule. As the data set is shuffled before the training and each molecule is provided with the same weight, a randomly nominated molecule is used to compare the closest and furthest neighbors from the generated data set. (B) A generated molecule with the highest similarity with the known molecule ($T_C \sim 0.82$). (C) A generated molecule with the highest dissimilarity form the known molecule ($T_C \sim 0.06$).

of the generated polymers for OPV applications by training a supervised regression model to predict electronic properties of these polymers. We consider a data set consisting of the optoelectronic properties of the OSCs based on polymer−fullerene constituent materials as a testbed. Figure 5A−D

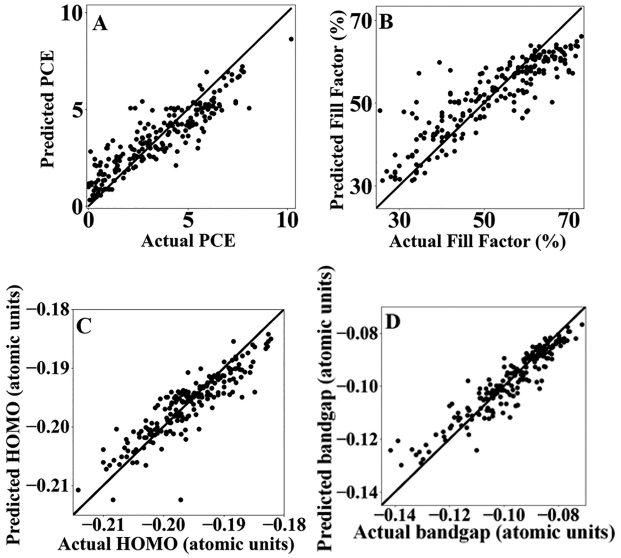


Figure 5. Predictability of random forest regressor for different properties of OPV devices. Comparison of the predicted against the actual properties for organic donor polymers from the known OPV data set. (A) Actual PCE values compared with the predicted PCEs. (B) Predicted FF relative to the actual FF. (C) Comparison between HOMO energies of the actual molecules and the corresponding predictions. (D) The predicted bandgap energy contrasted against the actual bandgaps. All these results indicate excellent predictive capability of the regression model.

illustrates the model training accuracy of the electrical properties such as PCE, FF, bandgap energy and HOMO energy levels of donor polymers from RF regressor model trained on 1400 OPV molecules. The physiochemical descriptors extracted from SMILES are standardized with zero mean and unit variance and accounted as the feature vector for the regression model. The OPV data set is divided into ∼85% as training and ∼15% as test data. Additionally, the RF model is optimized for the hyperparameters employing a random search approach with 5-fold cross validation. We consider mean squared error (MSE) as the performance metric to analyze the predictive capability. With the OPV data set, the MSE for the regression model is calculated as 1.007 (PCE), 11.200 (FF), 0.00001 (HOMO), and 0.00003 (bandgap), and the R^2 score is about 0.80 (PCE), 0.76 (FF), 0.88 (HOMO), and 0.88 (bandgap).

The R^2 score obtained from the fit suggests that the RF based regression provides an effective predictive model for the properties of the new polymer materials. On the basis of the trained regression model, we plot the response surface of the electrical properties such as the PCE and FF, as illustrated in Figure 6A and B. We note the predicted properties for the

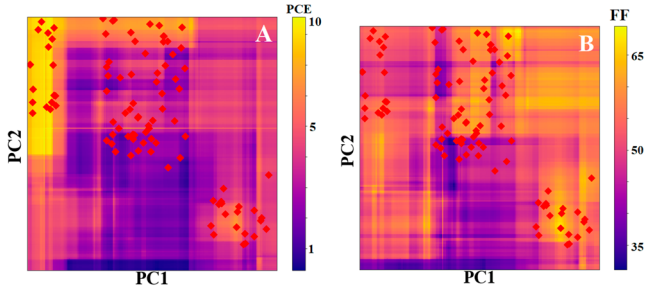


Figure 6. Performance evaluation of generated polymer repeat units. (A) The response surface plot for PCE along the two major PCA components from the extracted physiochemical descriptors. 100 new molecules (red diamonds) are chosen from different regions of the overall surface. At least 20% of the molecules are found to be in the high PCE region asserting their potential as OPV materials. (B) The response surface plot for FF along the major PCA components. Similar to the PCE, generated molecules are found in the hotspots with high FF.

generated OPV molecules on the response surface. Figure 6A shows that at least 20% of the valid molecules predict PCE > 10% and can be considered as potential candidate OPV donor materials. As PCE is correlated with FF, the predictions for these properties show similar trend. However, from Figure 6A and B, different regions of interest emerge which could be used to screen new polymer materials from different regions of the chemical space.

Figure 7 illustrates the chemical structure of the generated monomers as a function of the similarity coefficient. Figure 7A presents the known monomer unit with the highest PCE obtained from the OPV training data set. On the basis of the Tanimoto similarity analysis from the extracted physiochemical features, we note that the electrical properties rely significantly on the underlying chemical structure of the monomer units. For instance, from a visual comparison of the structures in Figure 7A−C, we find that the presence of thiophene rings contributes to high PCE and FF. This observation strongly agrees with the experimental literature demonstrating the remarkable improvement of device performance in the

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

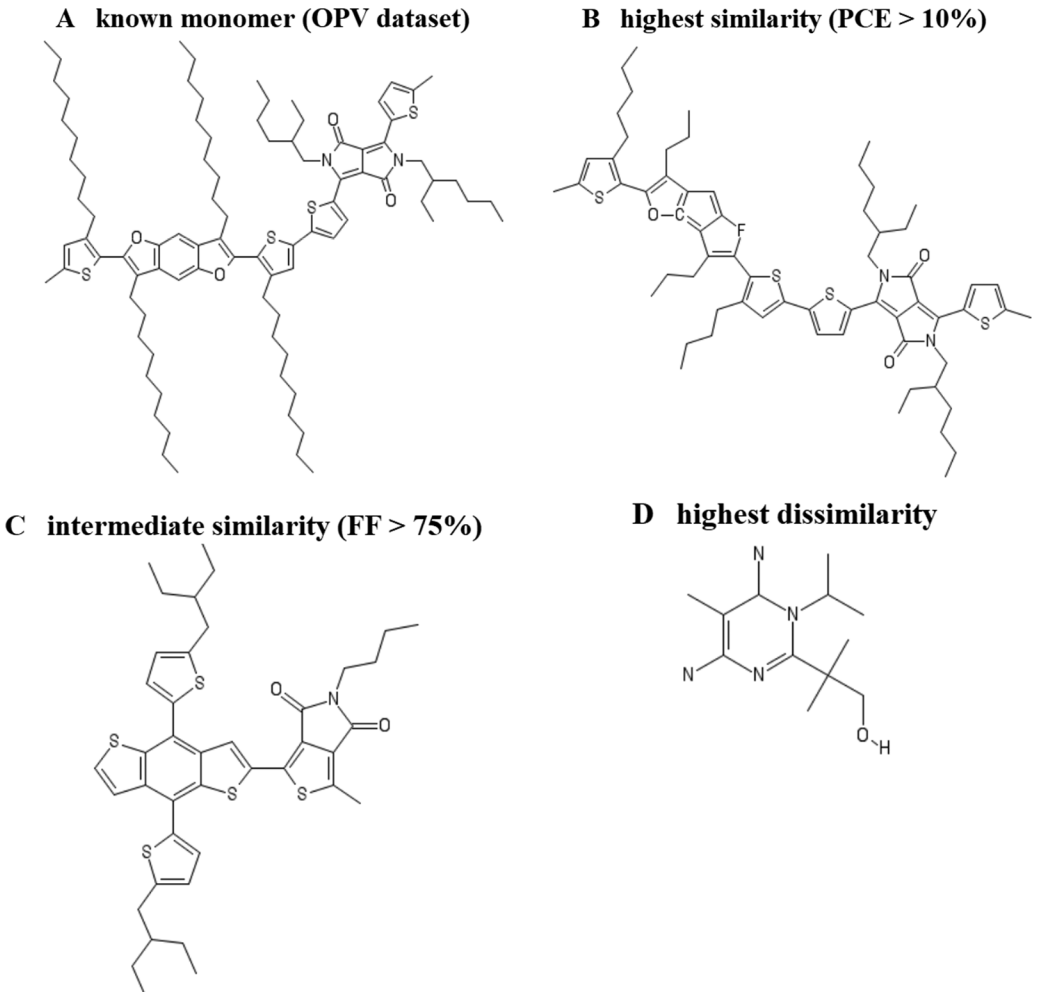


Figure 7. Comparison between transfer learned donor polymers and best performing donor from OPV data set. (A) The chemical structure of the known monomer unit from the OPV data set demonstrating highest PCE (~10.7%). (B) A generated monomer with highest similarity (T_C ~ 0.73). The molecular structure predicts ~10.9% power conversion efficiency from the RF-based regression model. (C) A generated monomer with highest FF (~79%). A moderate similarity (T_C ~ 0.50) with Figure 7A reveals correlation between PCE and FF with respect to the underlying chemical specificity of the donor material. (D) A generated monomer with the strongest dissimilarity (T_C ~ 0.05). The absence of thiophene rings in the repeat unit evince the poor performance because of the low PCE and FF.

presence of thiophene rings in bulk heterojunction (BHJ) active layers.[60,61]

CONCLUSION

In summary, we employ LSTM model to learn the grammar and vocabulary of one of the popular chemical languages known as the SMILES notation. As deep learning models are data intensive and typically a data set of more than 100 000 data points is necessary for the model to be robust, application of these approaches in materials design problems such as accelerated materials discovery is challenging. Inspired by the transfer learning in de novo drug design, we implement the approach to perfect a pretrained LSTM model trained on a small focused data set of OPV donor materials. Our results suggest that transfer learning enables ~90% valid molecule generation spanning the chemical space in contrast to the <25% valid molecule generation in absence of the pretrained network. An average Tanimoto similarity ~0.4225 ± 0.010 corroborates the potential of the predictive scheme to incorporate chemical information embedded in the underlying SMILES notation. Physio-chemical features extracted from the molecular fingerprint accurately predicts the performance metrics of OPV devices from the known database. The response surface plot on reduced principal component dimensions reveals the existence of generated molecules in the region of interest, promising a potential enhancement in PCE of OPV devices.

METHODS

All the LSTM models are implemented using TensorFlow GPU 2.1.0[62] and Keras 1.1.0[63] in Python 3.7.7. We leverage the power of GPU computing on a Dell precision tower with 2.2 GHz Intel Xeon E5 processors and Nvidia GeForce GTX 1080 GPU. Validation of SMILES string and extraction of physiochemical descriptors are performed using RDkit chemical library[58] implemented in Python. PCA and random forest regressions to predict optoelectronic properties from the physiochemical features are employed using Scikit-learn libraries.[64]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01157.

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142

GDB large data set (1 million small organic molecules), GDB small data set (50 000 validation/test organic molecule data set), and focused data set for OPV transfer learning (includes experimental data of measured PCE, FF, short circuit current, open circuit voltage, molecular weight, polydispersity index and calculated molecular orbital energy, and bandgap energy) (ZIP)

AUTHOR INFORMATION

Corresponding Author

Ganesh Balasubramanian − *Department of Mechanical Engineering & Mechanics, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0003-1834-5501; Phone: +1-610-758-3784; Email: bganesh@lehigh.edu

Authors

Joydeep Munshi − *Department of Mechanical Engineering & Mechanics, Lehigh University, Bethlehem, Pennsylvania 18015, United States*

Wei Chen − *Department of Mechanical Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0002-4653-7124

TeYu Chien − *Department of Physics & Astronomy, University of Wyoming, Laramie, Wyoming 82071, United States;* orcid.org/0000-0001-7133-6650

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01157

Author Contributions

J.M. performed the machine learning on the data and post processing of the results. J.M. and G.B. analyzed the results and wrote the manuscript. W.C. and T.C. assisted with the discussion of the results and editing the manuscript.

Notes

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors' and do not necessarily reflect the views of the NSF.

The authors declare no competing financial interest.

The data set required to reproduce the results presented in this work are provided as Supporting Information. The data generated during the LSTM training can be made available from the corresponding author upon reasonable request. The codes required to reproduce these findings cannot be shared at this time as it also forms part of an ongoing study. However, the results can be reproduced following the computational methods described in this paper.

ACKNOWLEDGMENTS

This material is based on the work supported by the National Science Foundation (NSF) under Award Nos. CMMI-1662435, 1662509, and 1753770 (the latter includes support from a supplement for data science related CMMI research activities).

REFERENCES

(1) Günes, S.; Neugebauer, H.; Sariciftci, N. S. Conjugated Polymer-Based Organic Solar Cells. *Chem. Rev.* **2007**, *107* (4), 1324−1338.

(2) Thompson, B. C.; Fréchet, J. M. J. Polymer−Fullerene Composite Solar Cells. *Angew. Chem., Int. Ed.* **2008**, *47* (1), 58−77.

(3) Chen, L.-M.; Hong, Z.; Li, G.; Yang, Y. Recent Progress in Polymer Solar Cells: Manipulation of Polymer: Fullerene Morphology and the Formation of Efficient Inverted Polymer Solar Cells. *Adv. Mater.* **2009**, *21* (14−15), 1434−1449.

(4) Scharber, M. C.; Sariciftci, N. S. Efficiency of bulk-heterojunction organic solar cells. *Prog. Polym. Sci.* **2013**, *38* (12), 1929−1940.

(5) Alessandri, R.; Uusitalo, J. J.; de Vries, A. H.; Havenith, R. W. A.; Marrink, S. J. Bulk Heterojunction Morphologies with Atomistic Resolution from Coarse-Grain Solvent Evaporation Simulations. *J. Am. Chem. Soc.* **2017**, *139* (10), 3697−3705.

(6) Carrillo, J.-M. Y.; Kumar, R.; Goswami, M.; Sumpter, B. G.; Brown, W. M. New insights into the dynamics and morphology of P3HT:PCBM active layers in bulk heterojunctions. *Phys. Chem. Chem. Phys.* **2013**, *15* (41), 17873−17882.

(7) Lee, C.-K.; Pao, C.-W.; Chu, C.-W. Multiscale molecular simulations of the nanoscale morphologies of P3HT:PCBM blends for bulk heterojunction organic photovoltaic cells. *Energy Environ. Sci.* **2011**, *4* (10), 4124−4132.

(8) Ghumman, U. F., Iyer, A.; Dulal, R.; Wang, A.; Munshi, J.; Chien, T.; Balasubramanian, G.; Chen, W. A Spectral Density Function Approach for Design of Organic Photovoltaic Cells in *ASME IDETC/CIE Design Automation Conference*; ASME: Quebec City, Quebec, Canada, 2018.

(9) Munshi, J.; Farooq Ghumman, U.; Iyer, A.; Dulal, R.; Chen, W.; Chien, T.; Balasubramanian, G. Composition and processing dependent miscibility of P3HT and PCBM in organic solar cells by coarse-grained molecular simulations. *Comput. Mater. Sci.* **2018**, *155*, 112−115.

(10) Munshi, J.; Dulal, R.; Chien, T.; Chen, W.; Balasubramanian, G. Solution Processing Dependent Bulk Heterojunction Nanomorphology of P3HT/PCBM Thin Films. *ACS Appl. Mater. Interfaces* **2019**, *11* (18), 17056−17067.

(11) Li, G.; Zhu, R.; Yang, Y. Polymer solar cells. *Nat. Photonics* **2012**, *6*, 153.

(12) Lee, M. R.; Eckert, R. D.; Forberich, K.; Dennler, G.; Brabec, C. J.; Gaudiana, R. A. Solar Power Wires Based on Organic Photovoltaic Materials. *Science* **2009**, *324* (5924), 232−235.

(13) Kaltenbrunner, M.; White, M. S.; Głowacki, E. D.; Sekitani, T.; Someya, T.; Sariciftci, N. S.; Bauer, S. Ultrathin and lightweight organic solar cells with high flexibility. *Nat. Commun.* **2012**, *3*, 770.

(14) Liu, Q.; Jiang, Y.; Jin, K.; Qin, J.; Xu, J.; Li, W.; Xiong, J.; Liu, J.; Xiao, Z.; Sun, K.; Yang, S.; Zhang, X.; Ding, L. 18% Efficiency organic solar cells. *Science Bulletin* **2020**, *65* (4), 272−275.

(15) Qin, J.; Zhang, L.; Xiao, Z.; Chen, S.; Sun, K.; Zang, Z.; Yi, C.; Yuan, Y.; Jin, Z.; Hao, F.; Cheng, Y.; Bao, Q.; Ding, L. Over 16% efficiency from thick-film organic solar cells. *Science Bulletin* **2020**, *65* (23), 1979−1982.

(16) Gertsen, A. S.; Castro, M. F.; Søndergaard, R. R.; Andreasen, J. W. Scalable fabrication of organic solar cells based on non-fullerene acceptors. *Flexible and Printed Electronics* **2020**, *5* (1), 014004.

(17) Qiu, Z.; Hammer, B. A. G.; Müllen, K. Conjugated polymers − Problems and promises. *Prog. Polym. Sci.* **2020**, *100*, 101179.

(18) Yu, G.; Gao, J.; Hummelen, J. C.; Wudl, F.; Heeger, A. J. Polymer Photovoltaic Cells: Enhanced Efficiencies via a Network of Internal Donor-Acceptor Heterojunctions. *Science* **1995**, *270* (5243), 1789−1791.

(19) Berger, P. R.; Kim, M. Polymer solar cells: P3HT:PCBM and beyond. *J. Renewable Sustainable Energy* **2018**, *10* (1), 013508.

(20) Marcon, V.; Raos, G. Molecular Modeling of Crystalline Oligothiophenes: Testing and Development of Improved Force Fields. *J. Phys. Chem. B* **2004**, *108* (46), 18053−18064.

(21) Wong-Ekkabut, J.; Baoukina, S.; Triampo, W.; Tang, I. M.; Tieleman, D. P.; Monticelli, L. Computer simulation study of fullerene translocation through lipid membranes. *Nat. Nanotechnol.* **2008**, *3*, 363.

(22) Moon, J. S.; Lee, J. K.; Cho, S.; Byun, J.; Heeger, A. J. Columnlike" Structure of the Cross-Sectional Morphology of Bulk Heterojunction Materials. *Nano Lett.* **2009**, *9* (1), 230−234.

(23) Peter, S.; Meyer, H.; Baschnagel, J. Molecular dynamics simulations of concentrated polymer solutions in thin film geometry.

II. Solvent evaporation near the glass transition. *J. Chem. Phys.* **2009**, *131* (1), 014903.
(24) Huang, D. M.; Faller, R.; Do, K.; Moulé, A. J. Coarse-Grained Computer Simulations of Polymer/Fullerene Bulk Heterojunctions for Organic Photovoltaic Applications. *J. Chem. Theory Comput.* **2010**, *6* (2), 526−537.
(25) To, T. T.; Adams, S. Modelling of P3HT:PCBM interface using coarse-grained forcefield derived from accurate atomistic forcefield. *Phys. Chem. Chem. Phys.* **2014**, *16* (10), 4653−4663.
(26) Root, S. E.; Savagatrup, S.; Pais, C. J.; Arya, G.; Lipomi, D. J. Predicting the Mechanical Properties of Organic Semiconductors Using Coarse-Grained Molecular Dynamics Simulations. *Macromolecules* **2016**, *49* (7), 2886−2894.
(27) Munshi, J.; Ghumman, U. F.; Iyer, A.; Dulal, R.; Chen, W.; Chien, T.; Balasubramanian, G. Effect of polydispersity on the bulk-heterojunction morphology of P3HT:PCBM solar cells. *J. Polym. Sci., Part B: Polym. Phys.* **2019**, *57* (14), 895−903.
(28) Munshi, J.; Chien, T.; Chen, W.; Balasubramanian, G. Elasto-morphology of P3HT:PCBM Bulk Heterojunction Organic Solar Cells. *Soft Matter* **2020**, *16*, 6743.
(29) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.
(30) Paul, A.; Jha, D.; Al-Bahrani, R.; Liao, W.; Choudhary, A.; Agrawal, A. Transfer Learning Using Ensemble Neural Networks for Organic Solar Cell Screening. *2019 International Joint Conference on Neural Networks (IJCNN)* **2019**, DOI: 10.1109/IJCNN.2019.8852446.
(31) Sahu, H.; Ma, H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *J. Phys. Chem. Lett.* **2019**, *10* (22), 7277−7284.
(32) 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* (2), 343−349.
(33) Sahu, H.; Yang, F.; Ye, X.; Ma, J.; Fang, W.; Ma, H. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J. Mater. Chem. A* **2019**, *7* (29), 17480−17488.
(34) 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. *Advanced Theory and Simulations* **2019**, *2* (1), 1800116.
(35) 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* (11), No. eaay4275.
(36) 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* (10), 2639−2646.
(37) Li, H.; Cui, Y.; Liu, Y.; Li, W.; Shi, Y.; Fang, C.; Li, H.; Gao, T.; Hu, L.; Lu, Y. Ensemble Learning for Overall Power Conversion Efficiency of the All-Organic Dye-Sensitized Solar Cells. *IEEE Access* **2018**, *6*, 34118−34126.
(38) Paul, A.; Furmanchuk, A.; Liao, W.-k.; Choudhary, A.; Agrawal, A. Property Prediction of Organic Donor Molecules for Photovoltaic Applications Using Extremely Randomized Trees. *Mol. Inf.* **2019**, *38* (11−12), 1900038.
(39) Jørgensen, P. B.; Schmidt, M. N.; Winther, O. Deep Generative Models for Molecular Science. *Mol. Inf.* **2018**, *37* (1−2), 1700133.
(40) Ruder, S.; Peters, M.E.; Swayamdipta, S.; Wolf, T. Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* **2019**, 15.
(41) Chary, M.; Parikh, S.; Manini, A.; Boyer, E.; Radeous, M. A Review of Natural Language Processing in Medical Education. *western journal of emergency medicine* **2018**, *20* (1), 78−86.
(42) Singh, P., Natural language processing. In *Machine Learning with PySpark*; Springer, 2019; pp 191−218.
(43) Eria, K.; Jayabalan, M. Neural Machine Translation: A Review of the Approaches. *J. Comput. Theor. Nanosci.* **2019**, *16* (8), 3596−3602.
(44) Lyu, C.; Chen, B.; Ren, Y.; Ji, D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinf.* **2017**, *18* (1), 462.
(45) Fu, Y.; Lou, F.; Meng, F.; Tian, Z.; Zhang, H.; Jiang, F. An intelligent network attack detection method based on rnn. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)* **2018**, 483.
(46) Tai, Y.; He, H.; Zhang, W.; Jia, Y. Automatic generation of review content in specific domain of social network based on RNN. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)* **2018**, 601.
(47) Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; Lei, M. Machine learning in materials science. *InfoMat* **2019**, *1* (3), 338−358.
(48) Hochreiter, S.; Schmidhuber, J.. LSTM can solve hard long time lag problems. In *Advances in Neural Information Processing Systems*, 1997.
(49) Liu, X. L. Deep Recurrent Neural Network for Protein Function Prediction from Sequence. *bioRxiv*, 2017, 103994. https://www.biorxiv.org/content/10.1101/103994v1.
(50) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, *53* (7), 1563−1575.
(51) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37* (1−2), 1700111.
(52) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120−131.
(53) Colby, S. M.; Nuñez, J. R.; Hodas, N. O.; Corley, C. D.; Renslow, R. R. Deep Learning to Generate in Silico Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples. *Anal. Chem.* **2020**, *92* (2), 1720−1729.
(54) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.
(55) 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), 160086.
(56) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Granada, Spain, 2011; pp 2546−2554.
(57) Bergstra, J., Yamins, D.; Cox, D. D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, Volume 28; JMLR.org: Atlanta, GA, USA, 2013; pp I−115−I−123.
(58) Landrum, G. *RDKit: Open-Source Cheminformatics*, 2006.
(59) Segaran, T. *Programming Collective Intelligence: Building Smart Web 2.0 Applications*; O'Reilly Media, Inc., 2007.
(60) Zhang, F.; Wu, D.; Xu, Y.; Feng, X. Thiophene-based conjugated oligomers for organic solar cells. *J. Mater. Chem.* **2011**, *21* (44), 17590−17600.
(61) Duan, C.; Gao, K.; Colberts, F. J. M.; Liu, F.; Meskers, S. C. J.; Wienk, M. M.; Janssen, R. A. J. Thiophene Rings Improve the Device Performance of Conjugated Polymers in Polymer Solar Cells with Thick Active Layers. *Adv. Energy Mater.* **2017**, *7* (19), 1700519.
(62) Abadi, M., Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D.G.; Steiner, B.;

Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: A system for large-scale machine learning in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*; USENIX Association: Savannah, GA, USA, 2016; pp 265−283.

(63) Chollet, F. *keras*, 2015.

(64) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

https://dx.doi.org/10.1021/acs.jcim.0c01157
J. Chem. Inf. Model. 2021, 61, 134−142