Hindawi
International Journal of Polymer Science
Volume 2019, Article ID 4538514, 7 pages
https://doi.org/10.1155/2019/4538514

Research Article

Virtual Screening of Conjugated Polymers for Organic Photovoltaic Devices Using Support Vector Machines and Ensemble Learning

Fang-Chung Chen [1,2]

[1]*Department of Photonics, National Chiao Tung University, Hsinchu 30010, Taiwan*
[2]*Center for Emergent Functional Matter Science, National Chiao Tung University, Hsinchu 30010, Taiwan*

Correspondence should be addressed to Fang-Chung Chen; fcchen@mail.nctu.edu.tw

Received 7 November 2018; Revised 25 February 2019; Accepted 3 March 2019; Published 31 March 2019

Academic Editor: Zhonghua Peng

Herein, we report virtual screening of potential semiconductor polymers for high-performance organic photovoltaic (OPV) devices using various machine learning algorithms. We particularly focus on support vector machine (SVM) and ensemble learning approaches. We found that the power conversion efficiencies of the device prepared with the polymer candidates can be predicted with their structure fingerprints as the only inputs. In other words, no preliminary knowledge about material properties was required. Additionally, the predictive performance could be further improved by "blending" the results of the SVM and random forest models. The resulting ensemble learning algorithm might open up a new opportunity for more precise, high-throughput virtual screening of conjugated polymers for OPV devices.

1. Introduction

Organic photovoltaic (OPV) devices have been attracting much attention because of their advantageous properties, including light weight, mechanical flexibility, low material and fabrication cost, and short energy payback times [1–4]. Apart from traditional solar panels, possible applications of OPV devices also include power generators for wearable electronics, portable devices, and the Internet of things (IoTs) [5–10]. The state-of-the-art OPV devices are prepared based on the concept of "bulk heterojunction." Although some novel nonfullerene acceptors have been identified as promising candidates [3], the photoactive layers of most efficient OPV devices reported so far consist of a conjugated polymer as the electron donor and a fullerene derivative as an electron acceptor, forming interpenetrating *p-n* heterojunction networks [1]. The high interfacial area between the donor and acceptor, which can overcome the high binding energy of excitons, ensures effective generation of free charge carriers upon solar irradiation, thereby resulting in a high photocurrent and a decent power conversion efficiency (PCE) [1–4].

The design of organic materials apparently is one of the focuses for OPV-related research. Unfortunately, the large size of chemical space, which is recently estimated at on the order of 10^6 molecules, makes rational material search very challenging [11–15]. For instance, the Harvard Clean Energy Project explored the molecular space through basic combination rules from an initial collection of 26 molecular fragments and resulted in calculation of material properties for 3.5 million materials [15]. Because the combination of the just 26 building blocks leads to such a great number of molecules for OPV applications, it required a distributed computing framework to implement such huge calculations [15]. Accordingly, there remains a need for an effective virtual screening method, which is capable to fast screen potential organic materials in parallel to the experimental selection and validation. In addition to high precision, an ideal screening platform also requires high level of generality and should be able to adapt itself to the rapid development of new materials.

Recently, machine learning (ML) has been extensively employed for accelerating the virtual screening of organic materials in various fields, such as organic light-emitting

diodes (OLEDs) and OPV devices [12–17]. In particular, multilayer perceptrons have been used to yield highly accurate predictions on many properties to accelerate material discovery for OPV devices [15]. More recently, Nagasawa et al. collected experimental results of more than 1200 conjugated polymers from ~500 literatures; the parameters included band gaps (E_g), molecular weights (M_w), energy levels, and fingerprints of the chemical structures [17]. The authors conducted supervised ML based on random forest (RF) and artificial neural network (ANN) models for material screening of potential polymer donors for bulk heterojunction OPV devices. Therefore, for the first time, they were able to design a conjugated polymer using machine learning algorithms.

The support vector machine (SVM) is one of the most common supervised machine learning algorithms [18]. It creates a hyperplane which separates the data into classes. The SVM is usually considered effective for datasets in which the number of feature dimensions is greater than the number of samples. Meanwhile, ensemble learning can improve the performance of machine learning through combining the outcomes from several models [19]. Aggregation of predictions of multiple models usually leads to better predictive performance compared to a single model. Indeed, ensemble methods have been widely used to solve many realistic problems. For example, the "Netflix prize" is aimed at improving the accuracy of predictions about how their customers rate a movie based on their previous preferences [20]. Many winners adopted ensemble learning in their recommendation engines, thereby achieving substantially improved performance.

Herein, we report results of ML for OPV applications using SVMs and compare the performance with the RF model. More importantly, the device PCEs were predicted only based on the fingerprints of the chemical structures. We found that the prediction performance of the models, which have been trained only with chemical fingerprints, was comparable with previous results [17]. In other words, the PCE values of the potential polymer candidates can be predicted without knowing any preliminary material properties. Further, the prediction accuracy of PCEs from the chemical structures was further improved by "blending" the results of SVM and RF models. The resulting ensemble learning algorithm might open up a new opportunity for more precise, high-throughput virtual screening of conjugated polymers for organic solar cells.

2. Experiment

The simplified molecular input line entry system (SMILES) codes and average PCE values were obtained from the previous results of Nagasawa et al. [17]. This dataset consists of 1203 polymers and the corresponding material and device properties. As illustrated in Figure 1, the chemical structures of the polymers were firstly converted to the repeating units. Then, the units were further transferred to SMILES codes. We used RDKit with python API to generate the chemical fingerprints from the SMILES codes [21]. There are many types of fingerprints to represent the chemical structures, including the molecular access system (MACCS) [17, 22] and extended connectivity fingerprint (ECFP) [17, 23]. In this work, we choose circular fingerprints, built by applying the Morgan algorithm, to set bits for the SMILES codes [21]. The number of bits was 2048 bits per hash. While converting the chemical structures to Morgan fingerprints, the radius of the fingerprint (r) is one of the important parameters, which takes the connectivity information into account. As the number "4" in ECFP4 corresponds to the diameter of the atom environments, Morgan fingerprints generated by the RDKit with a radius of 2 are roughly equivalent to ECFP4. During the ML processes, the whole dataset was split into training and testing subsets; 25% of the dataset was include in the testing split. Note that the same split subsets were used for different models in order to obtain consistent evaluation results. Figure 1 displays the typical process flow of the whole ML prediction. After obtaining the Morgan fingerprints from the original chemical structures, the training subset together with the corresponding PCE values was fitted into the models. Then, the model accuracies were evaluated by applying the testing subset and the corresponding PCE values.

3. Results and Discussion

We initially chose a Morgan radius of 2, which is similar to ECFP4, to generate chemical fingerprints. The radius of 2 indicates that the maximum diameter of the circular neighborhoods is 4 in the molecule. The number is usually sufficient for similarity searching. Figure 2(a) shows the initial results of the optimized SVM model while the radius of the Morgan fingerprint was set at 2. The validation of the model was performed with 301 samples (25% of the dataset). A correlation coefficient (R) was used to evaluate the model performance. The maximum R value is positive one, which indicates a perfect match between the predicted and experimental PCE values [17]. The R value of this model was 0.587. The major problem of the result was the overestimated PCE values for the samples with low experimental PCE; the PCE was also underestimated once the experimental PCE was larger than ~7%.

In order to further improve the accuracy, we mapped the effects of the Morgan radius on the model accuracies. We performed ten times simulation and obtained the correlation coefficient for each run. Figure 2(b) displays the influence of the Morgan radius on the performance of the SVM model. We could see that the accuracy depended slightly on how the dataset was split. The standard deviation (σ) of the ten runs was as small as 0.003 while the Morgan fingerprint was set at 2. On the other hand, the average R value was improved roughly with the increasing radius although we still observed some fluctuations (Figure 2(b)). In order the obtain a robust model, we arbitrarily selected the Morgan fingerprint of 5, which means that it considered 10 nearest atom neighborhoods in the molecules for the following works. As a result, the average R value of the SVM model achieved 0.633 ± 0.018.

After the determination of the parameters for the Morgan fingerprint, another ten times calculations were

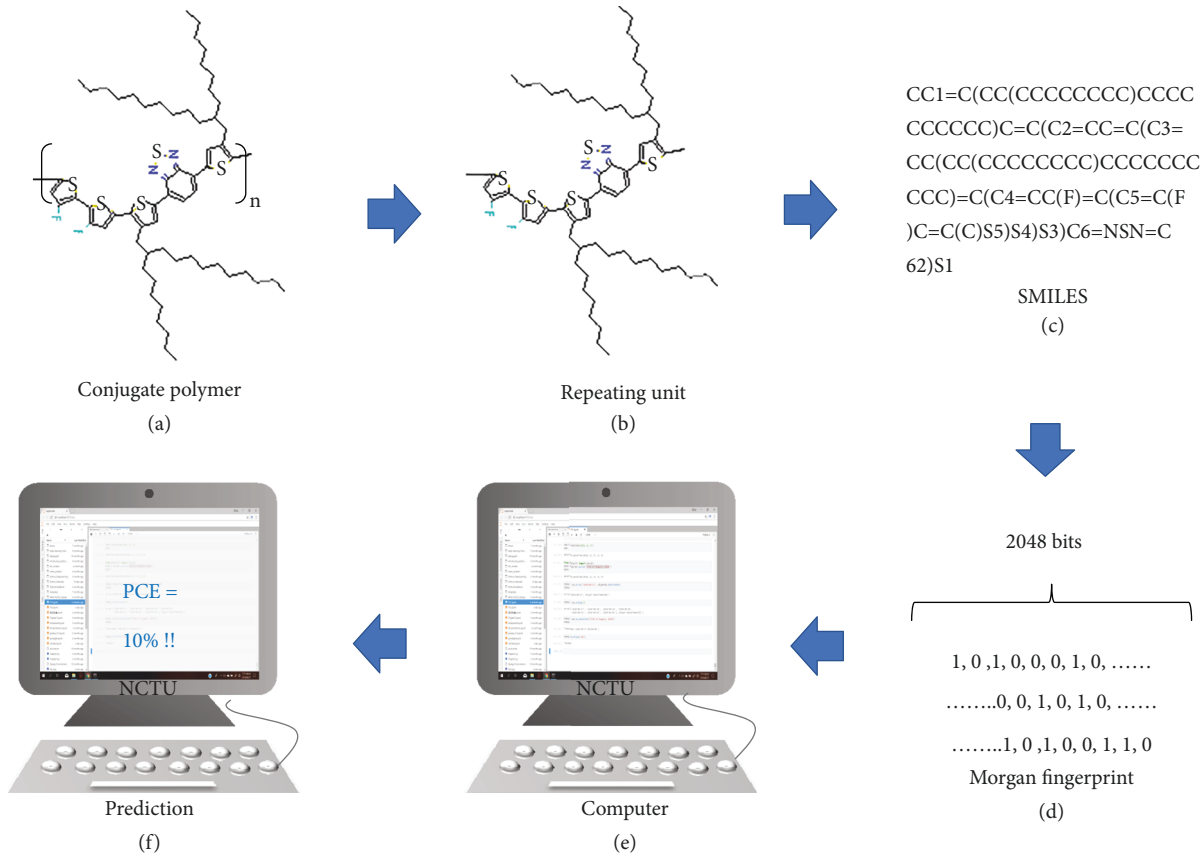

Figure 1: Schematic workflow of the machine learning for predicting the performance of conjugated polymers in OPV devices: (a) the chemical structure of the polymer; (b) the repeating unit; (c) the SMILES code; (d) the Morgan fingerprint; (e) the input of the dataset; (f) the resulting prediction.

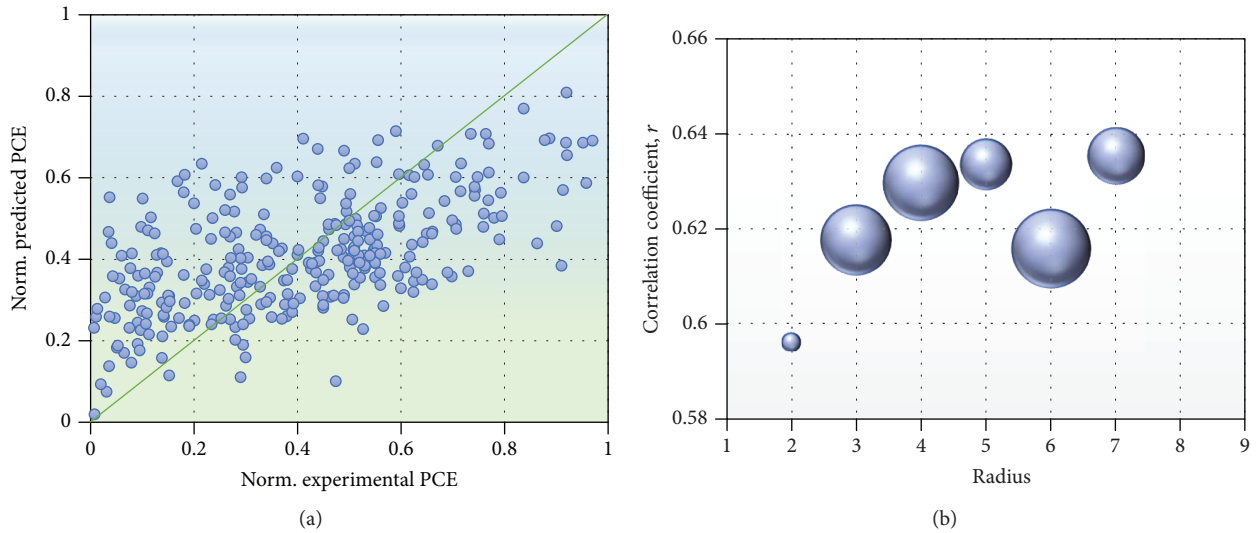

Figure 2: (a) The predictive result of the SVM model; the radius (r) of the Morgan fingerprint was two. Note that the experimental and predicted PCE values have been normalized. The diagonal line implies the positive perfect correlation ($R = 1$). (b) Effect of the Morgan radius on the performance of the SVM model. Note that the diameter of each bubble represents the relative size in quantity for the standard derivatives of the correlation coefficients.

performed; the same split dataset was used for other two models, which will be discussed later, for each run. Figure 3(a) displays a typical result of the SVM; the correlation coefficient was 0.627. This R value is comparable with that reported earlier, indicating acceptable results of the SVM model [17]. The performance of the SVM model was

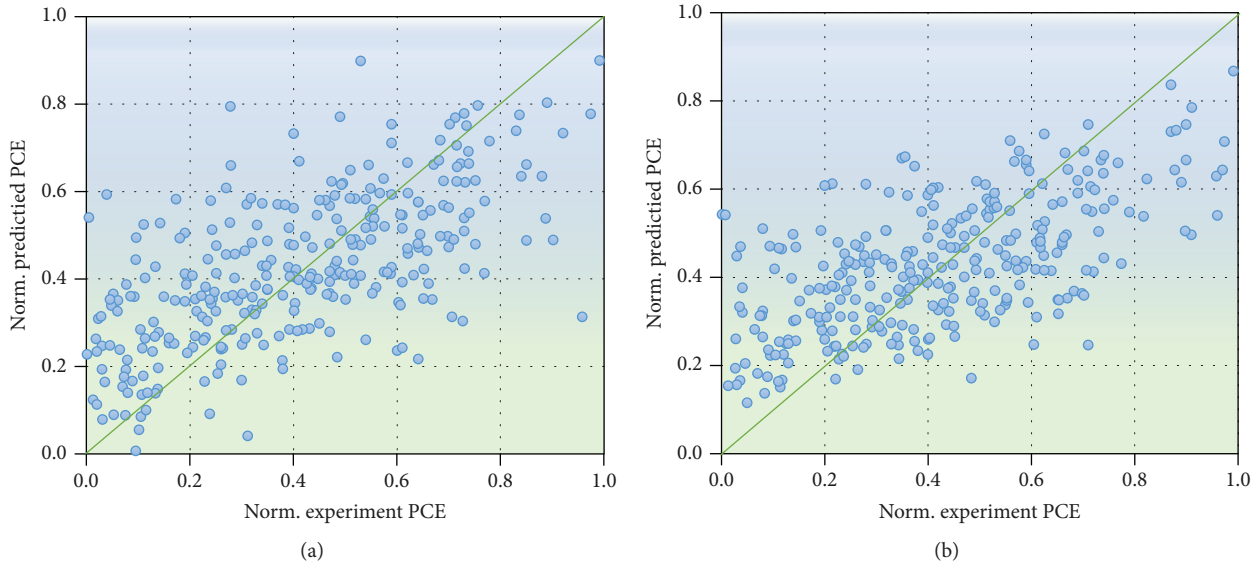

Figure 3: (a) The predictive result of the SVM model; the radius of the Morgan fingerprint was five. (b) The predictive result of the RF model using the same split dataset.

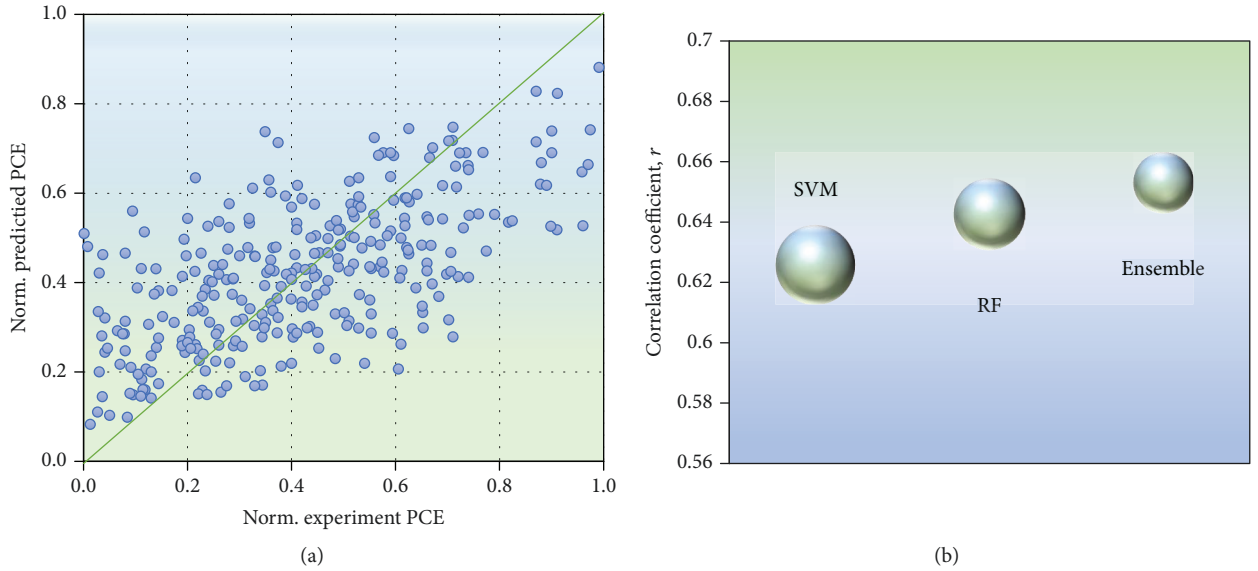

Figure 4: (a) The performance of the ensemble learning through averaging of the PCE values from the previous SVM and RF models. (b) Summary of the three models in this study. Note that the diameter of each bubble represents the relative size in quantity for the standard derivatives of the correlation coefficients.

also compared with that of the RF algorithm. Figure 3(b) revealed the result of the RF model applying the same split dataset; the R value was 0.640. It seems like that the RF model was slightly better than the SVM algorithm. One should note that the predication was only based on the information of SMILES of the polymers. In other words, we predicted the PCEs solely from the chemical structures of the polymers. Previously, Nagasawa et al. applied the digital keys, either MACCS (166 digital keys) or ECFP6 (1064 bits), together with the information about energy levels of the highest occupied molecular orbital (HOMO), E_g, and M_w of the polymers, to the RF model [17]. In our cases, we only adopted Morgan fingerprints with 2048 bits as the input and very similar performance was achieved. The results were indeed not surprising. The addition of the other information would only add 3 additional bits in our inputs, which only led to trivial effects on the prediction results. Therefore, our work suggests that one can possibly predict the PCE of a particular polymer even without any preliminary knowledge about material properties.

As we indicated in the introduction, ensemble learning is one way for improving the model with the current dataset. In fact, RF itself is one kind of ensemble learning; it builds numbers of decision trees and aggregates the results to obtain

Table 1: Summary of the three models in this work. Note that the same split dataset was used for model training and testing.

Model	SVM	RF	Ensemble
Correlation coefficient (R)	0.626 ± 0.020	0.643 ± 0.018	0.653 ± 0.015

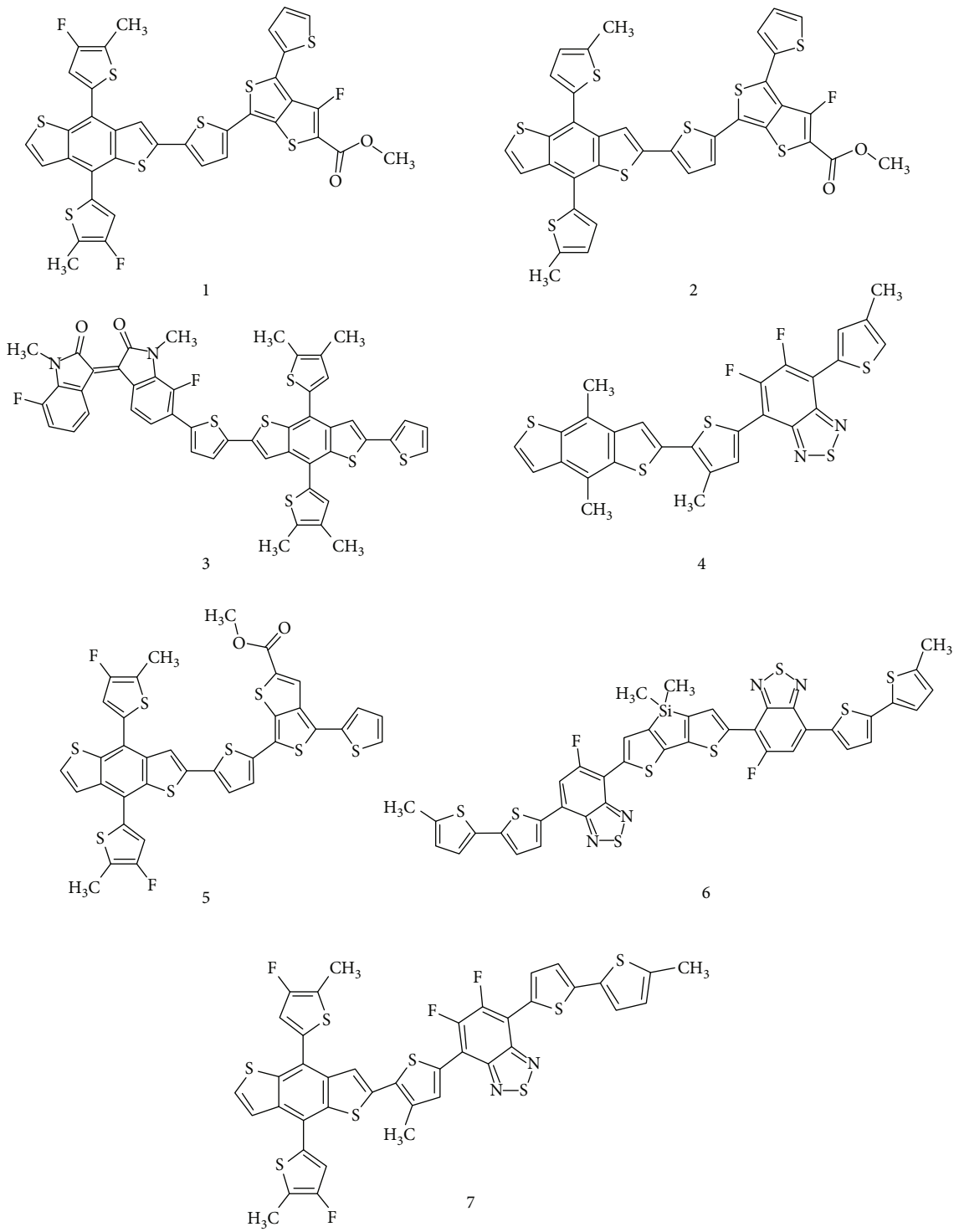

Figure 5: The chemical structures (1-6) of the repeating units that we predicted from the other dataset consisting of 316 compounds. Structure 7 is the one that we proposed; it was predicted to have the highest PCE value among the seven compounds in this figure.

a more accurate and stable prediction [17]. Because we have two models, SVM and RF, with similar prediction ability, we, therefore, considered to "blend" the predictions from the two models to increase predictive accuracy. We used the same training set to train the models individually and simply averaged the predictive PCE values of the same test set from the two models. The resulting normalized PCE values of the ensemble learning are illustrated in Figure 4(a). The typical performance of the ensemble learning was better than those of the SVM and RF models. Especially, the number of the data points extremely deviated from the ideal diagonal line ($R = 1$), implying perfect prediction, was reduced after the output data from the individual model was averaged. As a result, the R value from ten runs was increased to 0.653 ± 0.015, indicating that the predictive performance was indeed improved using such ensemble learning. The R values of the three models are summarized in Table 1.

In order to demonstrate the function of the model, we applied a dataset from the previous report, which contains 316 polymers [24]. Using the above ensemble model, we screened the compounds and selected six compounds with the highest PCE values; the repeating units are depicted in Figure 5. Note that we employed all the 1203 polymers in order to have the best performance. The order of the compounds followed the decreasing trend of the predictive PCE values. For instance, compound 1 was predicted to exhibit the highest PCE value; the second place was compound 2. From the structures, one common feature of these six compounds is the presence of fluoride atom(s). Fluorination of the conjugated polymer backbone has been considered as a promising approach for the development of high-performance polymers in OPV devices [25]. The addition of fluoride atoms would lower the HOMOs of the donor polymers, thereby increasing the open-circuit voltage. Moreover, fluorination also leads to the planarization of the backbone. The better morphology of the polymer blends possibly improves the charge transportation [25]. Therefore, OPVs prepared with these compounds should exhibit high PCEs. This is indeed a very important research trend over the past few years in this field.

Based on the chemical features from the above results, we propose one new structure (structure **7**). The predictive PCE value of the compound was even higher than that of compound 1 from the model. Further synthesis works should be done to confirm our prediction. The source code of the models as well as the prediction works can be found in reference [26].

In view of the current predictive results, apparently, the performance is still far from being satisfactory for virtual screening. From our current knowledge of OPV devices, the power conversion efficiency is very sensitive to the device processing conditions, including polymer purities, processing solvents [27], the additives [28, 29], the annealing methods [30, 31], device structures [32, 33], and interfacial materials [34]. The measurement conditions, for example, tested in a N_2-filled glove box or after encapsulation and the environment temperature, will also affect the value of PCEs. These variations make accurate prediction very difficult. Therefore, it is very important to standardize these conditions to improve the quality of the raw data. The other method is to setup one protocol to digitize these variations, which allows the machine to systematically learn the parameters. We expect that the predictive performance can be substantially improved once we improve the quality of the data and/or dramatically increase the number of data entries.

4. Conclusion

In this work, we demonstrated that the PCE values of the bulk-heterojunction OPV devices can be predicted just using the information about the chemical structure of the polymer donor in the device. A correlation coefficient higher than 0.60 could be obtained using SVM and RF models. The predictive performance was comparable with that of the RF algorithm using inputs considering other properties, such as band gaps and molecular weights. The results of these two models were further ensembled to generate more accurate predictions. Because the ML reported herein does not require huge calculation capability, we anticipate that such ensemble learning algorithm can pave a new avenue for high-throughput virtual screening with even higher prediction accuracies in the near future. Although the accuracies of the models revealed in this work are still far from being satisfactory for virtual screening, we believe that the results reported herein have already pushed the virtual screening of organic materials for solar applications one step further toward precise prediction and design of high-performance materials using artificial intelligence.

Data Availability

Previously reported dataset, consisting of 1203 polymers and the corresponding material and device properties, was used to support this study and is available at doi:10.1021/acs.jpclett.8b00635. These prior studies and datasets are cited at relevant places within the text in [17].

Conflicts of Interest

The author declares that he has no conflicts of interest.

Acknowledgments

The author thanks the Ministry of Science and Technology, Taiwan (grant nos. MOST 106-2221-E-009-127-MY3 and MOST 107-3017-F009-003), and the Ministry of Education, Taiwan (SPROUT Project-Center for Emergent Functional Matter Science of National Chiao Tung University), for the financial support. This work is also financially supported by the Research Team of Photonic Technologies and Intelligent Systems at NCTU within the framework of the Higher Education Sprout Project by the MOE in Taiwan. The author would also like to thank the Data Sciences Hack Week in the 233rd Electrochemical Society Meeting for inspiring this work.

References

[1] L. Lu, T. Zheng, Q. Wu, A. M. Schneider, D. Zhao, and L. Yu, "Recent advances in bulk heterojunction polymer solar cells," *Chemical Reviews*, vol. 115, no. 23, pp. 12666–12731, 2015.

[2] Q. An, F. Zhang, J. Zhang, W. Tang, Z. Deng, and B. Hu, "Versatile ternary organic solar cells: a critical review," *Energy & Environmental Science*, vol. 9, no. 2, pp. 281–322, 2016.

[3] J. Hou, O. Inganäs, R. H. Friend, and F. Gao, "Organic solar cells based on non-fullerene acceptors," *Nature Materials*, vol. 17, no. 2, pp. 119–128, 2018.

[4] J. Zhang, L. Zhu, and Z. Wei, "Toward over 15% power conversion efficiency for organic solar cells: current status and perspectives," *Small Methods*, vol. 1, no. 12, 2017.

[5] G. Dennler, S. Bereznev, D. Fichou et al., "A self-rechargeable and flexible polymer solar battery," *Solar Energy*, vol. 81, no. 8, pp. 947–957, 2007.

[6] C. L. Cutting, M. Bag, and D. Venkataraman, "Indoor light recycling: a new home for organic photovoltaics," *Journal of Materials Chemistry C*, vol. 4, no. 43, pp. 10367–10370, 2016.

[7] S.-S. Yang, Z.-C. Hsieh, M. L. Keshtov, G. D. Sharma, and F.-C. Chen, "Toward high-performance polymer photovoltaic devices for low-power indoor applications," *Solar RRL*, vol. 1, no. 12, 2017.

[8] N. W. Teng, S. S. Yang, and F. C. Chen, "Plasmonic-enhanced organic photovoltaic devices for low-power light applications," *IEEE Journal of Photovoltaics*, vol. 8, no. 3, pp. 752–756, 2018.

[9] D. Landerer, D. Bahro, H. Röhm et al., "Solar glasses: a case study on semitransparent organic solar cells for self-powered, smart, wearable devices," *Energy Technology*, vol. 5, no. 11, pp. 1936–1945, 2017.

[10] F.-C. Chen, "Emerging organic and organic/inorganic hybrid photovoltaic devices for specialty applications: low-level-lighting energy conversion and biomedical treatment," *Advanced Optical Materials*, vol. 7, no. 1, 2019.

[11] J.-L. Reymond, R. van Deursen, L. C. Blum, and L. Ruddigkeit, "Chemical space as a source for new drugs," *MedChemComm*, vol. 1, no. 1, pp. 30–38, 2010.

[12] Y. Zhu, R. Huang, R. Zhu, W. Xu, R. Zhu, and L. Cheng, "DeepScreen: an accurate, rapid, and anti-interference screening approach for nanoformulated medication by deep learning," *Advancement of Science*, vol. 5, no. 9, 2018.

[13] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud et al., "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[14] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel et al., "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," *Nature Materials*, vol. 15, no. 10, pp. 1120–1127, 2016.

[15] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the Harvard clean energy project: the use of neural networks to accelerate materials discovery," *Advanced Functional Materials*, vol. 25, no. 41, pp. 6495–6502, 2015.

[16] E. O. Pyzer-Knapp, G. N. Simm, and A. Aspuru Guzik, "A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials," *Materials Horizons*, vol. 3, no. 3, pp. 226–233, 2016.

[17] S. Nagasawa, E. Al-Naamani, and A. Saeki, "Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest," *The Journal of Physical Chemistry Letters*, vol. 9, no. 10, pp. 2639–2646, 2018.

[18] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.

[19] Y. Lan, Y. C. Soh, and G. B. Huang, "Ensemble of online sequential extreme learning machine," *Neurocomputing*, vol. 72, no. 13-15, pp. 3391–3395, 2009.

[20] October 2018, https://www.netflixprize.com/.

[21] October 2018, https://www.rdkit.org/.

[22] October 2018, http://rdkit.org/docs/source/rdkit.Chem.MACCSkeys.html.

[23] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[24] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm et al., "The Harvard Organic Photovoltaic Dataset," *Scientific Data*, vol. 3, 2016.

[25] N. Leclerc, P. Chávez, O. Ibraikulov, T. Heiser, and P. Lévêque, "Impact of backbone fluorination on π-conjugated polymers in organic photovoltaic devices: a review," *Polymers*, vol. 8, no. 1, 2016.

[26] https://github.com/FC-Richard-Chen/V-Screening-OPV.

[27] S. Zhang, L. Ye, H. Zhang, and J. Hou, "Green-solvent-processable organic solar cells," *Materials Today*, vol. 19, no. 9, pp. 533–543, 2016.

[28] X. Zhu, F. Zhang, Q. An et al., "Effect of solvent additive and ethanol treatment on the performance of PIDTDTQx:$PC_{71}BM$ polymer solar cells," *Solar Energy Materials and Solar Cells*, vol. 132, pp. 528–534, 2015.

[29] F. C. Chen, H. C. Tseng, and C. J. Ko, "Solvent mixtures for improving device efficiency of polymer photovoltaic devices," *Applied Physics Letters*, vol. 92, no. 10, 2008.

[30] G. Li, Y. Yao, H. Yang, V. Shrotriya, G. Yang, and Y. Yang, ""Solvent annealing" effect in polymer solar cells based on poly(3-hexylthiophene) and methanofullerenes," *Advanced Functional Materials*, vol. 17, no. 10, pp. 1636–1644, 2007.

[31] F. C. Chen, C. J. Ko, J. L. Wu, and W. C. Chen, "Morphological study of P3HT:PCBM blend films prepared through solvent annealing for solar cell applications," *Solar Energy Materials and Solar Cells*, vol. 94, no. 12, pp. 2426–2430, 2010.

[32] G. Li, C. W. Chu, V. Shrotriya, J. Huang, and Y. Yang, "Efficient inverted polymer solar cells," *Applied Physics Letters*, vol. 88, no. 25, 2006.

[33] F.-C. Chen, J.-L. Wu, and Y. Hung, "Spatial redistribution of the optical field intensity in inverted polymer solar cells," *Applied Physics Letters*, vol. 96, no. 19, 2010.

[34] Z. Yin, J. Wei, and Q. Zheng, "Interfacial materials for organic solar cells: recent advances and perspectives," *Advancement of Science*, vol. 3, no. 8, 2016.