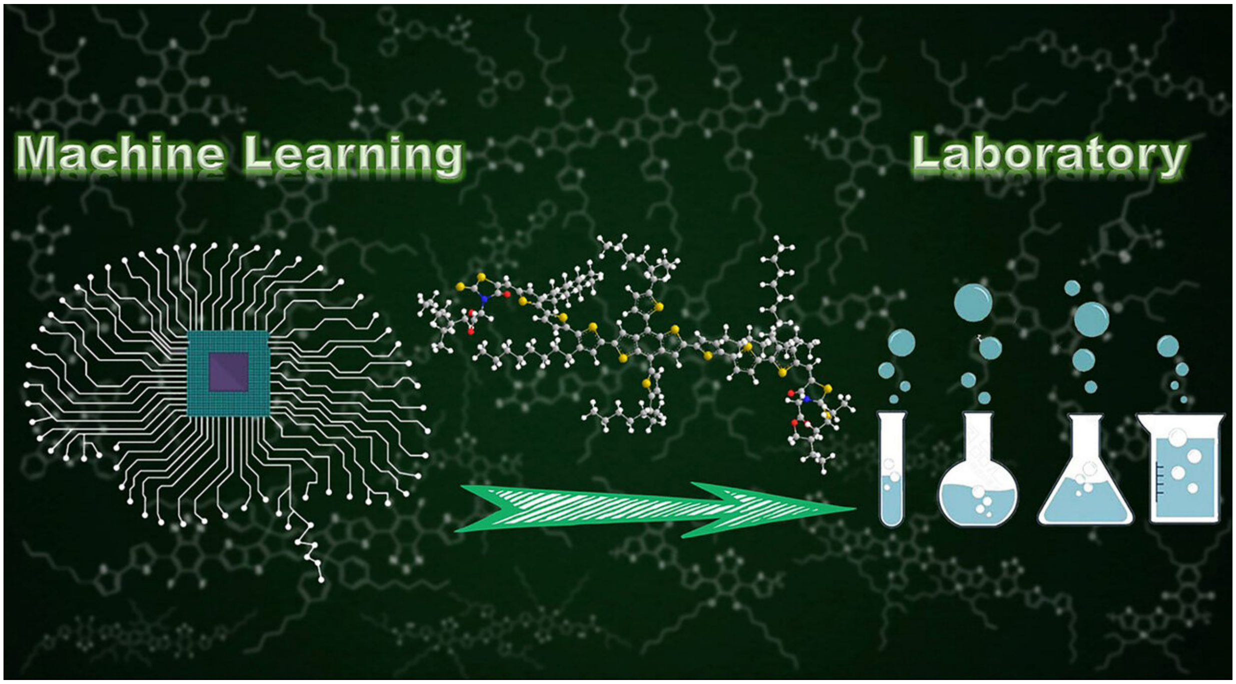
Machine learning-assisted molecular design for high-performance organic photovoltaic materials

November 19 2019, by Thamarasee Jeewandara



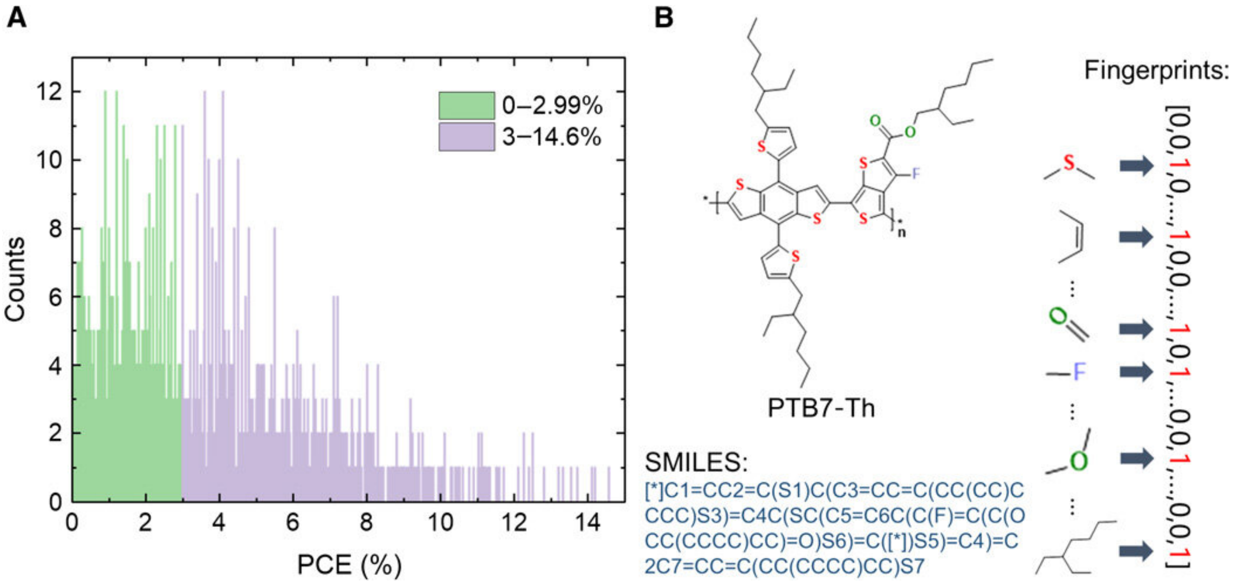
Using machine learning to assist molecular design. Credit: Wenbo Sun, Science Advances, doi: 10.1126/sciadv.aay4275

To synthesize high-performance materials for organic photovoltaics (OPVs) that convert solar radiation into direct current, materials scientists must meaningfully establish the relationship between chemical structures and their photovoltaic properties. In a new study on *Science*

Advances, Wenbo Sun and a team including researchers from the School of Energy and Power Engineering, School of Automation, Computer Science, Electrical Engineering and Green and Intelligent Technology, established a new database of more than 1,700 donor materials using existing literature reports. They used supervised learning with machine learning models to build structure-property relationships and fast screen OPV materials using a variety of inputs for different ML algorithms.

Using molecular fingerprints (encoding a structure of a molecule in binary bits) beyond a length of 1000 bits Sun et al. obtained high ML prediction accuracy. They verified the reliability of the approach by screening 10 newly designed donor materials for consistency between model predictions and experimental outcomes. The ML results presented a powerful tool to prescreen new OPV materials and accelerate the development of OPVs in materials engineering.

Organic photovoltaic (OPV) cells can facilitate direct and cost-effective transformation of solar energy into electricity with rapid recent growth to exceed power conversion efficiency (PCE) rates. Mainstream OPV research has focused on building a relationship between new OPV molecular structures and their photovoltaic properties. The traditional process typically involves the design and synthesis of photovoltaic materials for the assembly/optimization of photovoltaic cells. Such approaches result in time consuming research cycles that require delicate control of chemical synthesis and device fabrication, experimental steps and purification. The existing OPV development process is slow and inefficient with less than 2000 OPV donor molecules synthesized and tested so far. However, the data gathered from decades of research work are priceless, with potential values remaining to be fully explored to generate high-performance OPV materials.

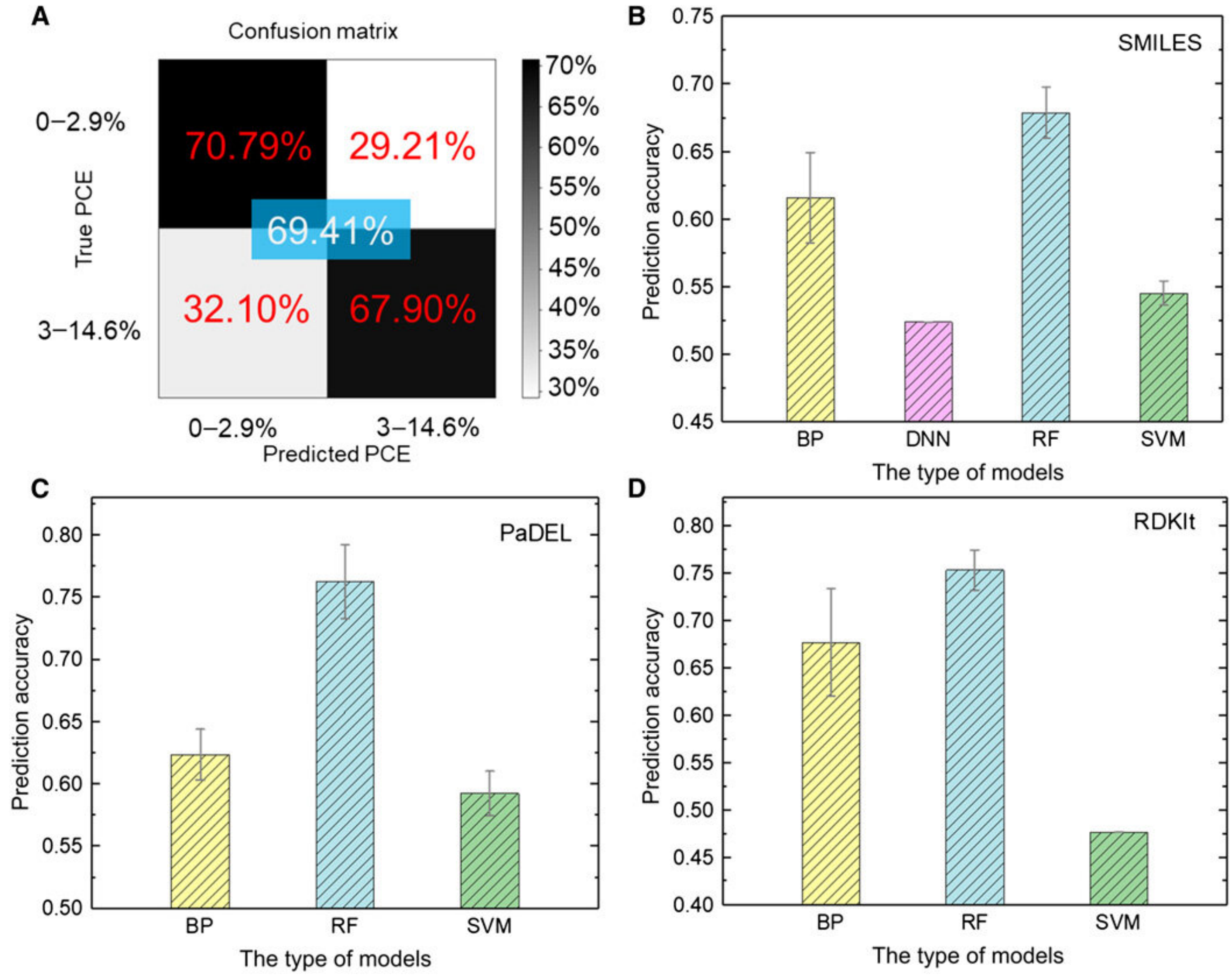


Information about the database of OPV donor materials. (A) Distribution of PCE values of the 1719 molecules in the database. (B) Schematics of expressions of a molecule, including image, simplified molecular-input line-entry system (SMILES), and fingerprints. Credit: Science Advances, doi: 10.1126/sciadv.aay4275

To extract useful information from the data, Sun et al. required a sophisticated program to scan through a large dataset and extract relationships from among the features. Since machine learning (ML) provides computational tools to learn and recognize patterns and relationships using a training dataset, the team used a data-driven approach to enable ML and predict diverse material properties. The ML algorithm did not have to understand the chemistry or physics behind the materials properties to accomplish the tasks. Similar methods have recently predicted the activity/properties of materials successfully during materials discovery, drug development and materials design. Prior to ML applications, scientists had generated cheminformatics to establish a useful toolbox.

Materials scientists have only recently explored the applications of ML in the OPV field. In the present work, Sun et al. established a database containing 1719 experimentally tested donor OPV materials gathered from literature. They studied the importance of programming language expression of the molecules first to understand ML performance. They then tested several different types of expressions including images, ASCII strings, two types of descriptors and seven types of molecular fingerprints. They observed the model predictions to be in good agreement with the experimental results. The scientists expect the new approach to greatly accelerate the development of new and highly efficient organic semiconducting materials for OPV research applications.

The research team first transformed the raw data into a machine readable representation. A variety of expressions exist for the same molecule comprising vastly different chemical information presented at different abstract levels. Using a set of ML models, Sun et al. explored diverse expressions of a molecule by comparing their predicted accuracy for power conversion efficiency (PCE) to obtain a deep-learning model accuracy of 69.41 percent. The relatively unsatisfactory performance was due to the small size of the database. For instance, previously when the same group used a larger number of molecules of up to 50,000, the accuracy of the deep-learning model exceeded 90 percent. To fully train a deep-learning model, researchers must implement a larger database containing millions of samples.

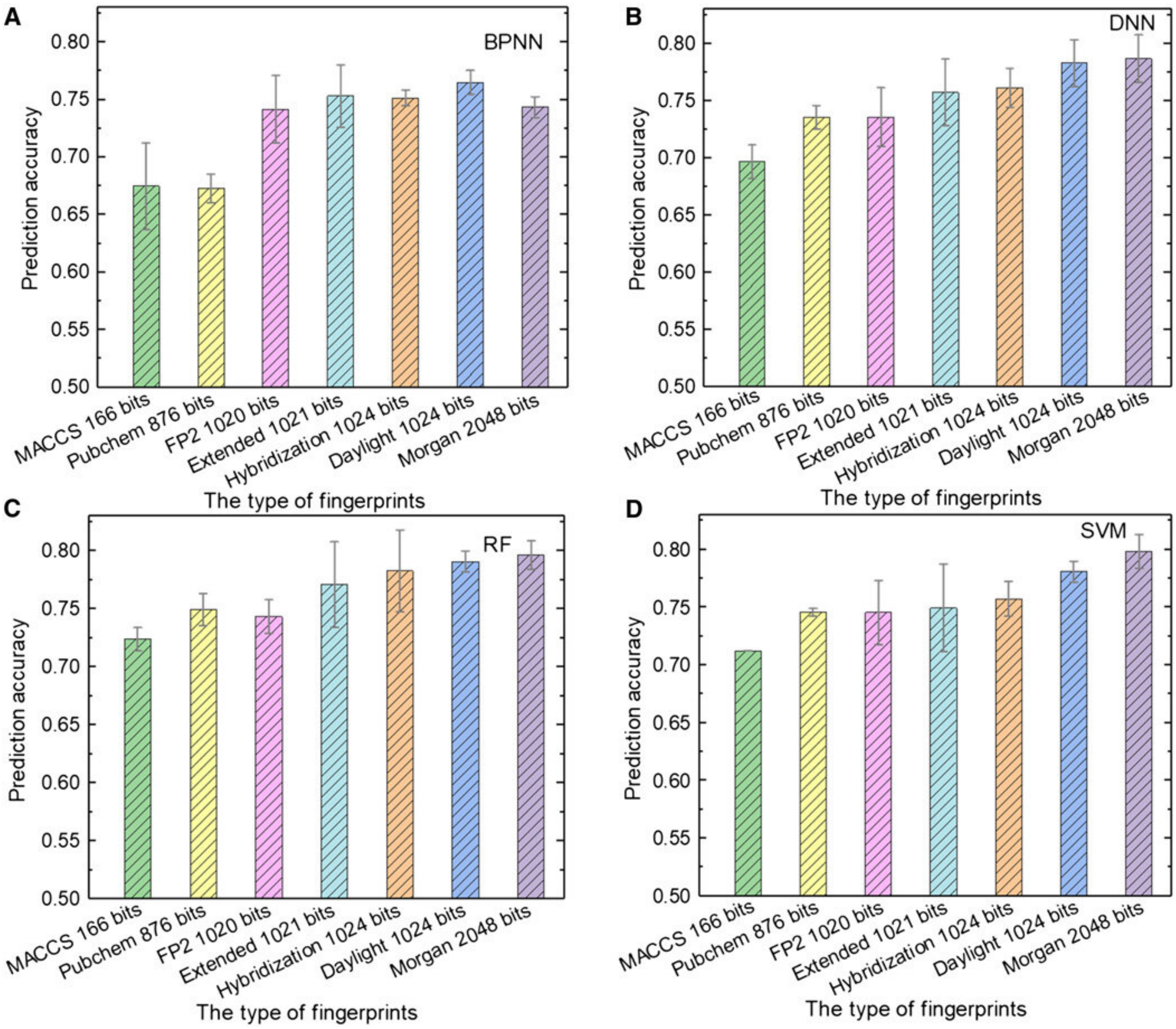


Testing results of ML models. (A) Testing of the deep learning model using images as input. (B to D) Testing results of different ML models using (B) SMILES, (C) PaDEL, and (D) RDKIt descriptors as input. Credit: Science Advances, doi: 10.1126/sciadv.aay4275

Sun et al. only had hundreds of molecules in each category at present, making it difficult for the model to extract enough information for higher accuracy. While it is possible to fine-tune a pre-trained model to reduce the amount of data required, thousands of samples are still necessary to accomplish a sufficient number of features. This led to the option of increasing the size of the database when using images to express molecules.

The scientists used five types of supervised ML algorithms in the study, including (1) back propagation (BP) neural network (BPNN), (2) deep neural network (DNN), (3) deep learning, (4) support vector machine (SVM) and (5) random forest (RF). These were advanced algorithms, where BPNN, DNN and deep learning were based on the artificial neutral network (ANN). The SMILES code (simplified molecular-input line entry system) provided another original expression of a molecule, which Sun et al. used as inputs for four models. Based on the results, the highest accuracy approximated 67.84 percent for the RF model. As before, unlike with deep learning, the four classical methods could not extract hidden features. As a whole, SMILES performed worse than images as descriptors of molecules to predict the PCE (power conversion efficiency) class in the data.

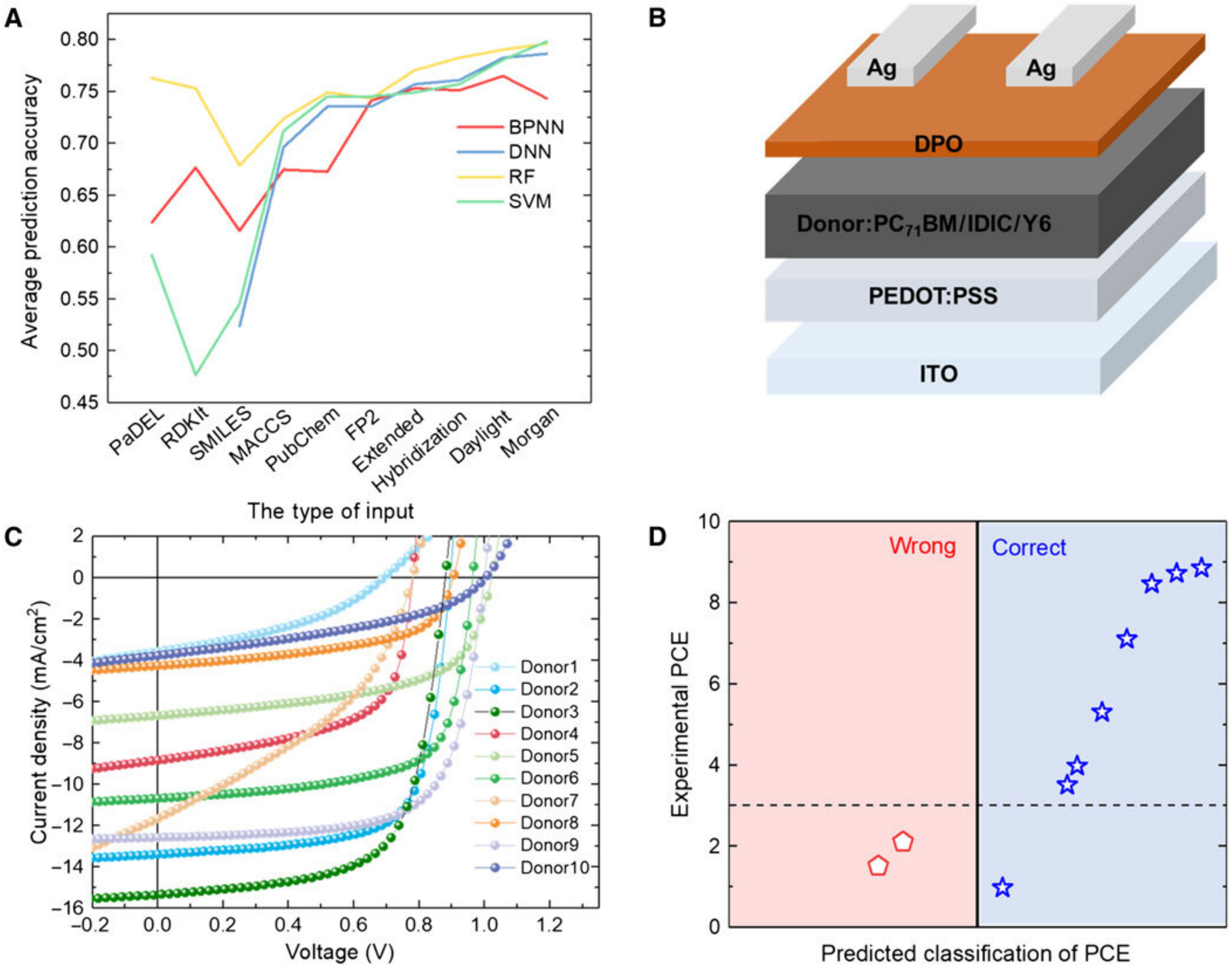
The researchers then used molecular descriptors that can describe the properties of a molecule using an array of numbers instead of the direct expression of a chemical structure. The research team used two types of descriptors PaDEL and RDKIt in the study. After extensive analyses across all ML models, a large data size implied more descriptors irrelevant to PCE affecting the ANN performance. Comparatively, a small data size implied inefficient chemical information to effectively train ML models, when using molecular descriptors as input in ML approaches, the key relied on finding appropriate descriptors that directly related to the target object.



Performance of ML models. (A to D) The testing results of (A) BPNN, (B) DNN, (C) RF, and (D) SVM using different types of fingerprints as input. Credit: Science Advances, doi: 10.1126/sciadv.aay4275.

The team next used molecular fingerprints; typically designed to represent molecules as mathematical objects and originally created to identify isomers. During large-scale database screening, the concept is represented as an array of bits containing "1" s and "0" s to describe the presence or absence of specific substructures or patterns within the molecules. Sun et al. used seven types of fingerprints as inputs to train the ML models and considered the influence of the fingerprint length on the prediction performance of different models to obtain diverse fingerprints. For instance, molecular access system (MACCS) fingerprints contained 166 bits and were the shortest input and the results were unsatisfactory due to their limited information.

Sun et al. showed the best combination of programming language and ML algorithm obtained using Hybridization fingerprints of 1024 bits and RF, to achieve a prediction accuracy of 81.76 percent; where Hybridization fingerprints represented SP2 hybridization states of molecules. When the fingerprint length increased from 166 to 1024 bits, the performance of all ML models improved since longer fingerprints included more chemical information.



Verification of ML models with experiment. (A) Comparison of the results from four different models. (B) Schematic diagram of the cell architecture used in this study. (C) J-V curve of the solar cell with the active layer using the predicted donor material. (D) Prediction results versus experimental data for the predicted donor materials with the RF algorithm and Daylight fingerprints. Credit: Science Advances, doi: 10.1126/sciadv.aay4275.

To test the reliability of the ML models, Sun et al. synthesized 10 new OPV donor molecules. Then used three representative fingerprints to express the chemical structure of the new molecules and compared the results predicted by the RF model and the experimental PCE values. The system classified eight of the 10 molecules. The results indicated the potential of the synthetic materials for OPV applications with additional experimental optimization for two of the new materials. A minor change in structure could cause a large difference in PCE values. Encouragingly, the ML models identified such minor modifications to facilitate favorable prediction results.

In this way, Wenbo Sun and colleagues used a literature database on OPV donor materials and a variety of programming language expressions (images, ASCII strings, descriptors and molecular fingerprints) to build ML models and predict the corresponding OPV PCE class. The team demonstrated a scheme to design OPV donor materials using ML approaches and experimental analysis. They prescreened a large number of donor materials using the ML model to identify leading candidates for synthesis and further experiments. The new work can speed up new donor material design to accelerate the development of high PCE OPVs. The use of ML in conjunction with experiments will progress materials discovery.

More information: Yann LeCun et al. Deep learning, *Nature* (2015). DOI: 10.1038/nature14539

Lingxian Meng et al. Organic and solution-processed tandem solar cells with 17.3% efficiency, *Science* (2018). DOI: 10.1126/science.aat2612

Wenbo Sun et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials, *Science Advances* (2019). DOI: 10.1126/sciadv.aay4275

Citation: Machine learning-assisted molecular design for high-performance organic photovoltaic

materials (2019, November 19) retrieved 23 April 2024 from
https://phys.org/news/2019-11-machine-learning-assisted-molecular-high-performance-photovoltaic.html