Performance Prediction and Experimental Optimization Assisted by Machine Learning for Organic Photovoltaics

Zhi-Wen Zhao, Yun Geng, Alessandro Troisi, and Haibo Ma*

The improvements of organic photovoltaics (OPVs) are mainly implemented by the design of novel materials and optimizations of experimental conditions through extensive trial-and-error experiments based on chemical intuition, which may be tedious and inefficient for exploring a larger chemical space. In the recent five years, data-driven methods using machine learning (ML) algorithms and the knowledge of known materials/experimental parameters are introduced to OPV studies to help build a quantitative structure-property relationship model and accelerate the molecular design and parameter optimization. Here, these recent promising progresses based on experimental OPV datasets are summarized. This review introduces the general workflow (e.g., dataset collection, feature engineering, ML model generation, and evaluation) of ML-OPV projects and discusses the applications of this framework for predicting OPV performance and experimental optimizations in OPVs. Finally, an outlook of future work directions in this exciting and quickly developing field is presented.

1. Introduction

Organic photovoltaic (OPV) based on π-conjugated molecules or polymers converts the solar energy to electricity utilizing the electron donor (D)/acceptor (A) bulk heterojunction structure to efficiently separate the bound excitons generated by light absorption into free charge carriers.[1] It has attracted immense research interest due to their numerous technology advantages such as inherent flexibility, portability, lightweight, low cost, and low toxicity,[2] as well as recently reported a high power conversion efficiency (PCE) of 19%.[3] However, to commercialize these technologies, further efficiency improvements are desired. This can be hardly accomplished by tedious and expensive trial-and-error methods based on chemical intuitions, labor-intensive synthesis, and characterization, especially when considering the large chemical space available for designing new organic donor or acceptor molecules and the numerous experimental parameters that can be tuned in the fabrication of the devices.[4]

It is important to realize that the construction of a quantitative structure–property relationship (QSPR) model[5] would greatly benefit the search of new functional molecules and the optimization of experimental conditions, and, accordingly, it can become a key factor to improve the photovoltaic performance of OPVs.[6] Furthermore, an interpretable QSPR model can also be used to unravel the fundamental working mechanism for a complex experimental behavior through its statistical correlation analysis with various microscopic physical parameters, which can hardly be achieved by blackbox predictions. Unfortunately, deriving such quantitative and interpretable QSPR models is highly nontrivial due to our current poor understanding of the complex photophysical phenomena in OPVs such as photoabsorption, exciton diffusion, exciton dissociation, charge generation and charge transport, as well as charge collection at electrodes.[1c,6] Earlier empirical QSPR models such as Scharber’s model[7] for fullerene-based OPVs and the modified Scharber’s model[8] for nonfullerene-based OPVs have been proposed for predicting OPV performance, yet, their accuracy for predicting PCE is unsatisfactory (with a linear Pearson correlation coefficient (r) between Scharber’s PCE and experimental PCE at around only 0.1–0.2).[9]

Driven by the development of artificial intelligence (AI)[10] algorithms, as well as the increased availability of a high-quality large dataset along with efficient data-mining techniques,
machine learning (ML) approaches have successfully resolved the difficulties of modeling the relationships between materials properties and complex chemical/physical factors. In recent years, ML techniques in conjunction with computational chemistry have been used to construct QSAR models to shed light on OPVs. Efficiency prediction accuracy has been raised to promising levels (e.g., \( r > 0.7 \)) using different ML algorithms either in fullerene or nonfullerene-based OPV devices. Some high-performing donor and/or acceptor molecule suggested by ML-assisted virtual screening have been successfully synthesized and experimentally verified. For example, Wu et al. synthesized three nonfullerene acceptors (NFAs) for PBDB-T-based OPVs with \( \approx 11\% \) experimental PCE, within 0.6% difference of the PCE predicted by random forest (RF) models.

Besides designing new molecules, the precise tuning of experimental conditions such as donor:acceptor (D:A) ratio, solvent additives, crystallinity, light type, etc. is equally crucial for the performance maximization of OPV devices. In addition, experimental verification for the predicted leading candidates is also restrained by these factors as they are highly related to syntheticsizability, cost, and stability. The stability and cost, as well as performance, could be considered as weighting factors for industrial figure of merit (i-FOM) during the manufacturing optimizations to facilitate commercialization. Moreover, essential factors such as synthetic complexity and cost of active layer (AL) donor polymers could be described quantitatively. Aforementioned progresses made it possible to use ML for discovering relationships between device performance and all relevant experimental parameters. For example, An et al. built an ML-OPV model by considering the deposition densities (DDs) as additional experimental descriptors based on a high-throughput fabrication dataset, thus successfully helping to find the optimal experimental condition for the roll-to-roll-processed PM6/Y6/IT-4F device to achieve the highest PCE at 10.2% to date.

In this review, first, we briefly present the ML workflow including the dataset generation, feature engineering, model construction, and evaluation in OPVs. Then, classifications and descriptions of ML descriptors are discussed in detail. Subsequently, we show examples of ML studies on predicting the performances of OPV molecules. In addition, how ML can be used for optimizing the device fabrication is discussed. Finally, we provide an outlook on ML perspectives for molecular explorations and experimental optimizations of OPVs.

2. General Workflow for ML-Assisted OPV Studies

The general workflow of ML studies to accelerate the discovery of new OPV molecules and the optimization of experimental conditions is shown in Figure 1, and the main steps are briefly introduced.

2.1. Dataset Generation

The definition of a valid dataset is a prerequisite for any data-driven methodology. To predict the macroscopic performance of OPV devices, ML-OPV studies usually rely on a dataset with sufficient experimental data, and, accordingly, the quality and size of this dataset affect the ML performance significantly. Generally, experimental datasets are somewhat biased (bias in reporting the best combination of materials and tendency to improve over the current best) and this creates difficulties for objectively evaluating how well the ML model will perform. Furthermore, the OPV dataset obtained from experimental results (Table 1) is usually much smaller than in other ML applications and the construction of data-driven models in OPV cannot be performed with routine methodologies but requires careful validations.

In the earlier stage of OPV research, most devices used fullerene acceptors (FAs) due to their advantages of high electron mobility and isotropic charge transport. Therefore, earlier ML works for OPV were usually based on FA-based datasets. Padula et al. and Sahu et al. built two FA datasets composed of 249 and 280 D/A pairs of small donor molecules, respectively. Later, Sahu et al. extended the small-molecule OPV dataset to around 300 data points. Experimental datasets for FA-based OPVs with polymer donors were also constructed by Nagasawa et al. with over 1,000 D/A pairs collected. At the same period, the ternary OPV dataset with 124 data points was constructed by Lee.

In recent years, NFAs have become a major focus of research in the development of OPVs. In contrast to the widely used FAs, the optical properties and electronic energy levels of NFAs can be easily tuned, leading to a rapid increase in PCEs for NFA-based OSCs, with values now exceeding 19%. Since 2019, a few NFA-based OPV datasets with data points around 100–600 have been built by several groups. Furthermore, Lee et al. and Hao et al. collected 135 and 157 experimental data points to construct ternary NFA-OPV datasets respectively.

To make the ML prediction applicable for general OPV systems, the diversity of FA- and NFA-based D/A pairs has to be accounted for in one dataset. Lopez et al. reported the Harvard Photovoltaic Dataset (HOPV15) containing 350 experimental data points with mixed FA- and NFA-based pairs. Later, Padula et al. and Zhao et al. constructed another mixed datasets with \( \approx 320 \) and 566 D/A pairs respectively, proposing a set of standard criteria for literature searches, in an effort to insure the reproducibility of the data collection and remove any possible bias in the definition of the dataset. Sun et al. constructed donor molecule-based experimental datasets with data points around 1,700–1,800 collected from the literature, in which the acceptor effect was neglected. These searches have been performed manually. Automatic searches attempted in different materials discovery areas proved difficult in this context. The difficulty in accessing experimental data also reflects in the digitalization of chemical compounds, for example, generating simplified molecular-input line-entry system (SMILES) from literature images is imperfect.

In a given D/A system, performances at different experimental conditions can be also collected to build datasets for ML training to optimize the experimental parameters. For instance, Du et al. considered ten processing parameters (D:A ratio, concentration, spin speed, solvents additives and their volume, annealing temperature and time for AL, material in electron transport layer [ETL], ETL annealing temperature/time) for over
100 processing parameter variations during the automated PM6/Y6 device fabrications.

2.2. Feature Selection

Features (descriptors) are the variables used in ML models, usually measurable via computational properties or characteristics of an observable phenomenon. The widely used features in ML-OPV studies include structural, electronic, and device parameters, which will be introduced in detail in Section 3. Feature selection methods have been adopted in OPV applications\textsuperscript{[21,22,27a,30]} owing to several reasons such as reducing the dimension and training time of the explored space, simplifying models, and improving accuracy.

The selection of the most informative, discriminating, and independent features minimizing the error rate by feature
selection algorithm is crucial for efficient ML tasks. In general, wrapper,[35] filter,[66] and embedded[37] methods are three main categories of algorithms scoring a feature subset or performing as part of the model construction process to propose new feature subsets with kinds of evaluation metrics. Accordingly, one of the most common approaches used in OPV field is the recursive feature elimination algorithm,[38] removing features with low weights and computing feature importance[39] for selecting the most informative ones. It can be noticed that some authors prefer not to perform any feature selection, but instead use intuition and the size of the dataset to set a fixed number of descriptors known for being important factors related to OPV performance.[14b,29]

2.3. Building QSPR Models by ML Algorithms

QSPR models can be built for predicting specific targets (e.g., OPV performance, stability, and synthesizability) in OPVs on the basis of using a reasonable dataset, descriptors, and suitable ML algorithms. A variety of ML methodologies have been applied for OPV. They have been presented in greater detail in numerous reviews[4c,12,13b,40] and textbooks, and[41] therefore their mathematical details are not repeated in this work. A group of methodologies including support vector regression (SVR),[42] support vector machine (SVM),[43] and kernel ridge regression (KRR)[44] can be described as an enhancement of conventional linear regression (LR)[45] methods to include nonlinear dependency of the descriptors and observables. Popular algorithms based on decision trees that have been used in OPV include RF,[46] and gradient boosting (GB).[47] The simplest possible algorithm, k-nearest neighbor (k-NN),[48] based on the distance between the neighbor configurations, has shown to be surprisingly effective in ML-OPV models. Conversely, artificial neural networks (ANN),[49] which contributed to the popularity of ML in the big-data project, are often less powerful in the presence of the dataset of limited size as common in OPV.

2.4. Assessment of Prediction Accuracy

For testing the prediction accuracy of ML algorithms, the dataset could be divided into the training set and testing set, some of which have additional external set for further validations. ML works can be performed with the leave-one-out (LOO), k-fold, or leave-one-cluster-out (LOCO)[50] cross-validation methods in search of the optimal hyperparameters of each algorithm. The prediction accuracy to the training/testing set is used to evaluate whether the built ML model is performing satisfactorily. Different papers report different measures and, sometimes, the algorithms are tuned to optimize different measures. Model performance metrics such as $r$, root-mean-square error (RMSE), mean square error (MSE), coefficient of determination ($R^2$), mean absolute percentage error (MAPE), and mean absolute error (MAE) can be adopted for regression models, which are defined as follows

$$\begin{align*}
    r &= \frac{\sum_{i=1}^{N} (R_i - \bar{R}) \times (P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N} (R_i - \bar{R})^2 \times \sum_{i=1}^{N} (P_i - \bar{P})^2}} \quad (1) \\
    \text{RMSE} &= \sqrt{\frac{\sum_{i=1}^{N} (R_i - P_i)^2}{N}} \quad (2) \\
    \text{MSE} &= \frac{\sum_{i=1}^{N} (R_i - P_i)^2}{N} \quad (3) \\
    R^2 &= 1 - \frac{\text{MSE}}{\text{var}(R_i)} \quad (4) \\
    \text{MAPE} &= \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i - R_i|}{|R_i|} \quad (5) \\
    \text{MAE} &= \frac{\sum_{i=1}^{N} |R_i - P_i|}{N} \quad (6)
\end{align*}$$

Here $N$ is the number of data points in the dataset; $R_i$ and $P_i$ represent the real and predicted values; $\bar{R}$ and $\bar{P}$ are the mean values for the real and predicted value, respectively. var($R_i$) is the variance of the sample data.

In the ML-OPV studies, some authors are interested in classifying candidates into well or bad performing sets instead of improving the accuracy of quantitative predictions. The classification accuracy is used to evaluate the performance of a classification model.[12a] Accuracy ($A$) is the ratio of making a correct classification, which can be described as

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

where $T$ and $F$ represent True and False, respectively, $N$ and $P$ represent negative and positive, respectively. Hence, $TP$ is the number of positive samples properly classified. Similarly, $TN$ is the number of negative samples that are classified accurately. The term $FP$ represents the number of negative samples that are classified as positive, while $FN$ is the number of positive samples that are classified as negative. In OPV research, $FP$ (it takes an experiment to discover) could be tolerated, while the $FN$ is likely going to remain undetected and it is a lost opportunity. Conversely, the error rate (ER = 1 − $A$) represents ratio of incorrect classifications.

Moreover, the precision ($P$) represents the fraction of positive classes that are classified as positive and the recall ($R$) is the fraction of positive classes that are actually positive, both of which are alternative types for accuracy in binary classification. They are often averaged in a single metric, the $F_1$ score (the harmonic mean of precision and recall), which is another indicator of the classification accuracy. Furthermore, false positive rate (FPR) is the ratio of negative classes that are classified as positive.

$$\begin{align*}
    P &= \frac{TP}{TP + FP} \quad (8) \\
    R &= \frac{TP}{TP + FN} \quad (9) \\
    F_1 &= \frac{P \cdot R}{P + R} \quad (10) \\
    \text{FPR} &= \frac{FP}{TN + FP} \quad (11)
\end{align*}$$
In addition, the receiver operating characteristic (ROC) curve\textsuperscript{[51]} plots the relationship between R and FPR. The area under the curve (AUC) of ROC can be interpreted as the proportion of correctly classified samples. Hence, the closer the AUC value is to one, the more the reasonable classification model is considered.

3. Description of Main Features

The features that are used to build ML methods are vastly different in terms of accuracy, cost, and availability and their choice influences both the accuracy of the method and its ability to be applied to a very large number of instances. We present them in this section grouped into three classes corresponding to different information they relay (Table 2).

3.1. Topological Information

The topological information from the molecule itself (molecular weight (W)/volume (V), number of $\pi$-electrons in donors ($N_{\text{elec}}$), number of unsaturated atoms in the main conjugation path ($N_{\text{atom}}$, etc.) is a kind of feature without any experimental measurements or computational evaluations, which could be quickly and cheaply obtained. Therefore, it plays an important role in exploring large chemical spaces.

In many cases, the SMILES of molecules are adopted as structural descriptors, which describe the structural species by short ASCII strings with letters and symbols, containing information on atoms, bonds, rings, aromaticity, and branches of a molecule.\textsuperscript{[52]} The development of cheminformatics, which has shown great success in the drug discovery using molecular fingerprints, has also greatly promoted ML-assisted material discovery.\textsuperscript{[53]} The fingerprinting procedures associating a structural fingerprint with each compound have been developed and tested in drug discovery. Similarity measures have been widely adopted in drug discovery with very low-cost methods (e.g., similarity-based regressions).\textsuperscript{[94]} The accurate and easily accessible programming language expressions provide possibilities for building a high-accuracy ML model without any computational data which need complicated calculations.

There is a comparatively smaller variation in functional groups in organic electronic materials with respect to drug-like molecules; hence, there is an opportunity of developing more specific fingerprints in the future. For a specific ML-OPV analysis, the ASCII strings can be converted to fingerprints with an array of bits with the same length.\textsuperscript{[14c]} It has been reported that the fingerprints with lengths above 1,000 bits contain substantial chemical information for ML-OPV predictions.\textsuperscript{[14c]} Simple “1-hot” binary vectors as indicator variables encode the acceptor molecules as 1 if present and 0 if absent in the ML model induced in the literature. Signature descriptors (a fingerprint whose elements indicate the number of each type of fragment that exists in each molecule) adopted in the ML studies can be generated for numerous materials within a short time and could generate fingerprints containing numerous fragments.\textsuperscript{[54]}

By pointing out the contributions of certain fragments to concrete properties, synthesis of new molecules or targeted modifications to existing materials become more approachable and fruitful for material scientists.

3.2. Electronic Structure Descriptors

More descriptors such as the electronic structure properties which do not depend only on the “shape” and (local) nonbonded interactions are considered in ML-OPV analysis going beyond topological descriptors. These electronic structure descriptors are directly influencing the microscopic optoelectronic processes in OPV and accordingly their quantitative performances. For instance, the simplest electronic structure parameters include energy levels of highest occupied molecular orbitals (HOMO) and lowest unoccupied molecular orbitals (LUMO), dipole moment ($\mu$), and vertical ionization potential (IP($\nu$)), which can be easily obtained from a ground-state quantum chemical (QC) calculation.\textsuperscript{[52,53]} It is well known that the ability of photo-absorption is affected by the wavelengths ($\lambda_{\text{abs}}$) and the associated oscillator strength ($f$) of the photosensitizer molecules’ excited states. To further consider the effect of detailed excited-state processes, the excited-state QC calculations can be carried out to provide other descriptors such as the parameters of $\lambda_{\text{em}}$ and $f$ we just mentioned and also the properties of other dark intermediate states (charge transfer ones and triplets)\textsuperscript{[56,21]} Parameters that affect the charge generation and charge dynamics such as hole–electron binding energy ($\beta_{\text{bind}}$), the energetic difference of LUMO of donor and acceptor ($E_{\text{LUMO}}$), reorganization energy (f), etc. can be widely used as additional descriptors, which could be obtained from QC calculations.\textsuperscript{[56]} Polarizability ($\beta$) is strongly correlated with molecular $\lambda$\textsuperscript{[51]} and there is a significant correlation between $\lambda$ and molecular size.\textsuperscript{[56]} The corresponding calculations normally require a few central processing unit (CPU) hours and so they are suitable to deal with the dataset of $\approx 10^2$–$10^5$ molecules but they cannot be extended to explore very large number of hypothetical molecules.

3.3. Macroscopic Properties

Experimental parameters can also be used to develop data-driven models for OPV. Examples of available parameters during experimental setups are concentration, solvents, additives, D:A ratio, annealing temperature/time, thickness, spin speed, and root-mean-square roughness (RMS) of atomic force microscope images. In addition, synthetic complexity\textsuperscript{[8]} and scalability factor\textsuperscript{[54]} could quantitatively characterize the synthetic cost, targeting the synthesizability of novel organic chemistries, which are useful in screening high-performing materials with low cost and easy synthesis. However, some important physical quantities in OPVs are challenging to characterize directly by experiments, such as solubility/miscibility and domain sizes, which play key roles in the film-forming process. It has been found that topology fingerprint encode very well properties related to lipophilicity (a feature related to miscibility, defined as the partition coefficient between n-octanol and water\textsuperscript{[57]}, which are also crucial for OPV.\textsuperscript{[58]} Also, recent works indicated that solubility can be predicted by ML methods with cheap and easily accessible factors\textsuperscript{[58]} (e.g., Hansen solubility parameters [HSP] could be
predicted through ML models by molecular size, electrostatics, charge density, and structural information\(^{[27a,59]}\). The aforementioned experimental parameters/properties, either obtained from direct measurements or theoretical modeling, are all adopted as descriptors for building a reliable QSPR model in the field of OPV.
4. ML-Assisted Performance Predictions of OPV Molecules

In this section, the recent research advances in ML-assisted performance predictions on OPV performance will be briefly overviewed. As mentioned in Section 2, the experimental OPV datasets are usually sorted into two main categories according to the types of the acceptor molecules. Therefore, here we first summarize different ML works based on FA- and NFA-based datasets respectively; then, recent ML predictions based on unified FA-/NFA-based datasets are introduced; subsequently, we discuss high-throughput virtual screening (HTVS) works based on the above ML models; finally, several experimental verifications are summarized.

4.1. Constructing QSPR Models with Different Types of Acceptors

4.1.1. FA-Based Dataset

Sahu et al.\cite{21} identified 13 important structural and electronic structure descriptors (Figure 2) to describe an FA-based dataset composed of 280 small donor molecules by an in-depth understanding of the microscopic mechanism of OPVs. Among them, one is the structural descriptor (number of unsaturated atoms), while others, such as polarizability, vertical ionization potential, and hole–electron binding energy, are related to the ground- and excited-state properties obtained by QC calculations. A range of ML algorithms including k-NN, RF, GB, and ANN were used to construct regression models for predicting PCE. Both tree-based models obtained remarkable prediction power with \( r \) at 0.79 and 0.76, respectively, as shown in Figure 2a,b. The importance of descriptors shown in Figure 2c,d suggests that the hole–electron binding energies are the most informative descriptors for GB and RF models. In addition to improving the overall performance of OPV, it is also necessary to meet other requirements for purpose-specific devices, such as high open-circuit voltage (\( V_{\text{OC}} \)) for solar–fuel energy conversion and high short-circuit current density (\( J_{\text{SC}} \)) for solar window applications. A follow-up work\cite{22} constructed ML models to predict other three important device parameters, \( V_{\text{OC}}, J_{\text{SC}}, \) and fill factor (FF). On the basis of the extended experimental dataset of 300 donor molecules, the prediction accuracy (\( r \)) of the GB model for \( V_{\text{OC}}, J_{\text{SC}}, \) and FF reached 0.67, 0.66, and 0.71, respectively.

Padula et al.\cite{9a} explored the possibility of using fewer physical descriptors but including more structural fingerprint descriptors in ML models based on an FA-based dataset composed of 249 D/A pairs with small donor molecules. Utilizing the Tanimoto similarity index between donors and Euclidean distance of electronic properties, they achieved high accuracies for the ML predictions of PCE. KRR and k-NN algorithms yield good prediction results (Figure 3) using both topological and electronic descriptors (\( r = 0.68 \) and 0.61 respectively). At the same time, Morgan fingerprints (\( r = 0.68 \)) were found to perform better than Daylight fingerprints (\( r = 0.59 \)) in KRR models.

Chen\cite{60} reported an ML study based on the FA-based dataset composed of 1,000 experimental parameters for polymer-fullerene-based OPVs\cite{23} using SVM and RF algorithms with cheap descriptors. Results indicated that the PCE values of OPV devices are predicted by the chemical information of polymer donors as the only inputs with \( r > 0.60 \). In addition,

---

Figure 2. a) Theoretically predicted versus experimental PCE for the testing set (30 molecules). b) All data points using the LOO crossvalidation technique for the GB model. c) The descriptor importance for the GB. d) The descriptor importance for the RF. Descriptors are in the following order: 1) \( N_{\text{atom}} \), 2) polarizability, 3) \( \Delta \gamma \), 4) \( \Delta \lambda \), 5) \( \Delta \beta \), 6) \( \Delta \zeta \), 7) \( \Delta \eta \), 8) \( \Delta \kappa \), 9) \( \Delta \lambda \), 10) \( \Delta \mu \), 11) \( \Delta \nu \), 12) \( \Delta \omega \), and 13) \( \Delta \chi \). Reproduced with permission.\cite{21} Copyright 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.
considering the promising ternary blended structure technology for OPV (incorporating third component in D/A blend) but the great difficulty in exploring the huge combination space of three components, Lee reported an ideal model of application of ML methods for ternary OPVs, by neglecting the synergistic effect. A dataset of 124 experimental devices and molecular energy levels was constructed to build ML models for ternary OPVs by adopting RF, GB, k-NN, LR, and ANN algorithms. It is demonstrated that RF models can also perform well for these ternary OPVs with $r \approx 0.77$ between the predicted and real PCEs.

### 4.1.2. NFA-Based Dataset

A major challenge in the construction of data-driven models for NFA is that they are often tested in combination with a variety of donors, thus complicating the analysis. A convenient approach is that attempted by Mahmood et al. who built a dataset of 283 experimental solar cells containing only the most common donor poly(3-hexylthiophene) (P3HT). As shown in Figure 4, both classification and regression ML algorithms were implemented with 3,000 molecular descriptors generated from the online chemical dataset. The classification accuracy of k-NN, SVM, RF, and ANN reaches 0.86, 0.88, 0.87, and 0.89, respectively. The regression analysis for PCE predictions suggests that SVM is the best model independent of the number of selected descriptors, whereas LR model is the best one for HOMO and LUMO energy-level predictions.

Wu et al. proposed a new approach to encode the chemical structure of polymers; hence, they were able to perform prediction of PCE for polymer/NFA heterojunction. The RF and boosted regression trees (BRT) showed better prediction accuracy with $r$ at 0.70 and 0.71 and RMSE at 1.17 and 2.42, respectively. Recently, Wen et al. constructed QSPR models with molecular descriptors and parameters related to morphology (D:A weight ratio and RMS roughness of atomic force microscope images) for NFA-based OPVs with more polymers chosen as donors. As shown in Figure 5, the constructed ML voting model (linear combination of the predictions from several regression models) on the basis of the engineering of structural, electronic, and device descriptors for a dataset containing 351 D/A pairs showed the best prediction accuracy ($r > 0.8$). The adopted experimental descriptors in this study are useful to improve the accuracy; however, it is an ongoing challenge to use ML to design new materials. Additional optimization works have to be implemented to get reasonable values for these experimental descriptors in new materials. For example, they conducted theoretically inverse optimizations of device specifications (RMS and D:A weight ratio) for screened D/A pairs, which provide feedback to experiments.

Some authors have successfully applied the previously developed model for FA to NFA like Lee with RF approach to predict the PCE for ternary blends achieving $R^2 > 0.80$. Another possibility is to focus on predicting some more specialized device characteristic which is likely to be easier to model. In this vain, Malhotra et al. used RF, GB, SVR, and ANN algorithms to predict nonradiative voltage loss. The GB algorithm with HOMO + LUMO + $E_g$ + RDKit/MACCS descriptors delivers overall better results ($r \approx 0.86$). Only molecular fingerprints as descriptors could produce good results ($r = 0.78$ and 0.73 for GB and SVR) as well.

### 4.1.3. Mixing of FA- and NFA-Based Dataset

The most well-known OPV dataset containing both FAs and NFAs is the HOPV15 dataset containing data for 350 experimental solar cells, mostly (273) fullerene based, with a greater proportion of polymeric donors (220) and bulk heterojunction experiments (270). In 2019, Paul et al. applied extremely randomized tree learning models to train on HOPV15 to predict HOMO energy levels for donor molecules with structural information as inputs, and they got high accuracy for prediction (96MAE of 1.91% and 1.97%). Meftahi et al. used the Bayesian regularized ANN with Laplacian prior (BRANNL) algorithm to train on HOPV15 dataset with signature descriptors for donors and 1-hot descriptors (1 means interested motif existing in the molecule, 0 as absent) for acceptors. The suggested descriptors have the advantage of being chemically interpretable building blocks that could be easily accessible. Their results suggest reasonable prediction accuracy for photovoltaic properties. Moreover, the key motifs that contribute to potential properties have been identified by sparse feature selection, which is consistent with microscopic simulations. It was concluded that the choice of descriptors is much more important than the selection of ML algorithms.

Padula and Zhao et al. also constructed unified FA-/NFA-based datasets with $\approx 320$ and 566 D/A pairs...
Figure 4. Brief description of the whole study. a) Model training. b) Library generation and screening of potential candidates. Reproduced with permission. Copyright 2021, the Royal Society of Chemistry.
respectively, considering only bulk heterojunction and molecular donor and acceptor. Especially the latter work contains a greater portion of NFA, about 82% of the total, reflecting the greater importance of NFAs in recent years. Padula et al.\textsuperscript{29} used several ML models (KRR, Gaussian processes regression [GPR],\textsuperscript{63} SVR, and $k$-NN) for the OPV performance predictions considering electronic properties and structural information, which produced impressive accuracy ($r = 0.78$). The reliability of the proposed model was verified by recently reported high-performance D/A pairs. Some of the screened promising D/A candidates in the dataset have been reported, again proving the reliability of predictions. The results suggest that more precise models could be obtained by widening the variety of experimental data. Zhao et al.\textsuperscript{30} explored the effect of increasing the number of descriptors and the size of datasets on ML studies for predicting efficiencies of all-small-molecule OPVs. Descriptors used in this work could be classified into structural (fingerprints) and physical ones (energy levels, molecular size, light absorption, and mixing properties). Results indicate that a larger dataset can enhance ML prediction power to achieve a high prediction accuracy ($r = 0.73$). Moreover, the addition of some descriptors (excited-state and miscibility properties) cannot significantly improve the prediction accuracy as they have been encoded in structural information, leading to minimal/redundant contribution.

According to the discussions of the above three subsections, it can be found that, once the dataset is sufficiently large, diverse, and homogeneous (of the order of a few hundred experimental data points), a variety of ML algorithms can produce QSPR models with predictive accuracies within a broadly similar range ($0.60 < r < 0.85$) despite the very different selections of the descriptors.

4.2. Designing Novel Chemical Materials or D/A Pairs Using ML Models

4.2.1. Screening Novel Chemical Materials/Combinations

To identify new promising OPV materials, it is desirable to explore the vast chemical space for appropriate organic semiconductor molecules, combining existing and new fragments with the almost limitless possibility of further functionalization. To this end, HTVS provides an effective approach to search for potential candidates for a given target property. Through screening a large number of compounds using computer-driven

Figure 5. Workflow of screening and optimization of D/A pairs and device parameters, including data collection of NFA-based OPVs, feature engineering of structural/electronic/device descriptors, constructing QSPR model, high-throughput screening of D/A pairs, searching device specifications to maximize PCE for the screened D/A pairs, and feedback to the experiments. Reproduced with permission.\textsuperscript{[13a]} Copyright 2021, American Chemical Society.
methods, invaluable insights into the identification and optimization of leading molecules can be obtained, thus triggering a meaningful interplay between computations and experiments. In 2014, the HOMO and LUMO values of 2.3 million donor molecules were studied in the Harvard Clean Energy Project (CEP) through QC calculations. Moreover, 51,000 NFAs were designed by 107 fragments, and their HOMO, LUMO energies, HOMO–LUMO gaps, computed $S_0 \rightarrow S_1$ transitions in the range of visible light (1.4–2.2 eV), and the oscillator strength were investigated for top candidates with predicted PCE larger than 9.0%. If descriptors in ML models are only topological, the number of compounds that can be computed is almost limitless because the evaluation is extremely cheap and the problem is only the generation of the sample materials. However, an accurate ML model requires QC calculations and so the number of molecules that can be tested is of the order of thousands. In this way, the field of HTVS contributes to create a dataset of hypothetical molecules to be evaluated by ML methods.

Recently, the well-trained ML models combined with HTVS have been implemented to accelerate the detection of OPV molecules. Meanwhile, generative models which are trained with molecular structures and their corresponding properties to generate new molecules with desired properties have been adopted for novel chemical compounds. This is referred to as a kind of inverse design aiming at discovering ideal molecules starting with desired functionality and ending in chemical space. In ML-OPV studies, researchers have generated novel polymer-based electron donors and NFAs by generative models. The following examples like the work reported by Sahu et al. were more combinatorial (combining potential building blocks in many possible ways). They reported HTVS of 10,170 donor molecules building from 32 fragments and 10 possible arrangements of donors, π-spacers, end-capping units, and acceptors. 1,000 lead candidates were screened out by a gradient boosting regression tree (GBRT) model with predicted PCE larger than 7.5%. The importance of building blocks was identified by Z-scores; as a result, benzothiophene (BDT), dithieno-benzodithiophene, and napho-dithiophene are identified as promising building blocks. Moreover, 126 promising candidates were screened out with a predicted PCE larger than 8% by ANN and GBRT models. Sun et al. identified 15 important building blocks from 1,758 different donor molecules using the RF model with in-house-designed fingerprint descriptors, as depicted in Figure 6. Through screening 18,960 donor candidates, they discovered that 20 molecules have good OPV performances when paired with Y6 acceptor. 6,337 molecules in the virtual material library were identified as promising donor candidates with predicted PCE > 8% by the GBRT model, among which 20 candidates were selected to construct OPV devices with Y6 acceptor. They built a Y6-based OPV dataset containing 44 OPV devices aimed at producing a suitable ML model for Y6-based OPVs. The ML results generated by GBRT yield acceptable prediction accuracy with RMSE of 2.20 and r of 0.74. There are 5 out of the 20 donor/Y6 pairs having predicted PCE larger than 15% (the largest PCE in the training dataset was 15.58%).

As seen in Figure 4, Mahmood et al. designed >3,000 small molecules of NFAs using various building blocks and 87 acceptors with predicted PCE > 7.5%. These molecules from a number of known fragments in OPVs were screened first by LR regression for HOMO and LUMO values and second by SVM regression for predicted PCE values. RF was chosen among alternative ML models to predict HSPs for the screened NFA molecule and identify the best “green” solvent for them. This example illustrates how ML can contribute in different ways to distinct aspects of the OPV design.

These HTVS studies considering the synergistic effect of D/A pairs have been reported recently. Wen et al. generated almost 2 million potential D/A pairs constructed from 9,963 donors and 194 experimental NFAs (Figure 5). It should be noted that the descriptors are distinct for D and A; thus, the computational cost scales as the sum of the number of donor and acceptor molecules instead of their product. Only three acceptors with structural similarity to Y6 molecule were found in these 1,501 D/A pairs whose predicted efficiencies are larger than 14%. The performance could be further optimized through tuning experimental conditions. For instance, the predicted PCEs of one of the pairs are in the range from 13.75% to 16.25%, with the best D:A ratio <0.8 and RMS ranging from 0.8 to 1.7 nm, respectively.

4.2.2. Experimental Verifications for Novel D/A Pairs

In recent years, there have been a few successful experimental corroborations of the new suggested OPV molecules by ML models, hence showing predictive power of the ML-guided rational molecular design.

Nagasawa et al. preliminarily selected 1,000 molecules with predicted PCE > 10% by Scharber’s model from the website of CEP, which could be used as donors in OPVs. Then, 149 molecules were generated with predicted PCEs in the highest region from the dataset of 1,000 molecules assisted by the supervised ANN and RF algorithms. Subsequently, they chose one molecule for experimental validation after considering synthesizability and π-conjugation. However, the experimental PCE (0.52%) mismatched the predicted ones (5.0–5.8% for different side chains); the failure may be caused in part by the poor performance of their ML model and the neglect of processing parameters.

Lin et al. predicted the efficiency of BO2FC8/m-ITIC-OR-4Cl pair by the RF algorithm, and the predicted result (11.2%) was found to almost perfectly match the experimental one (11.0%). At the same time, Sun et al. trained ML models (backpropagation [BP] neural network [BPNN], deep neural network [DNN], deep learning, SVM, and RF, as seen in Figure 7a) through images and ASCII strings on a dataset with 1,719 OPV donor materials to classify the materials into “low” and “high” performance (PCE < 3% were regarded as “low” and PCE > 3% was regarded as “high”). Ten designed donor molecules were chosen for sandwiched OPV devices as experimental validations (Figure 7b,c). As shown in Figure 7d, eight molecules were proper characterized into the correct class, meaning that the predicted PCEs are consistent with its experimental values with minor differences, indicating that with proper methodologies and solid datasets favorable ML-assisted predictions are feasible.

In 2020, the ML-assisted polymer/NFA combination screening for experimental validations has been reported by Wu et al. As demonstrated in Figure 8, over 32 million new D/A combinations were generated based on the structural fragments of polymers and NFAs in the dataset. Six new D/A
combinations (with PM6 and PBDB-T as donor material) with high predicted PCE by RF and BRT models and easy synthesis were selected for experimental validations. Encouragingly, five out of six real PCEs resemble to the predicted ones. Specifically, predicted efficiencies of two out of three devices (13.18% and 15.71% predicted by RF model) are relatively close to experimental ones (10.52% and 13.33%) for PM6-based systems, but experimental and the predicted PCEs using the RF model of PM6/Y-ThCH3 are not at the same level (6.67% vs. 10.41%). As for PBDB-T-based systems, the predicted PCEs using RF model (11.49%, 11.64%, and 11.32%) are all close to the experimental ones (11.02%, 11.08%, and 11.19%). The deviation between the predicted and experimental value is inevitable as PCEs of OPV materials are sensitive to the processing conditions, materials purity, external environment, etc.

Hao et al.\cite{27d} tested the reliability of ML models by experimental characterizations for the ternary OPV dataset. Their RF model is able to correctly classify PM6/Y6/IT-M combinations into the high class with PCE > 16% but however fails to distinguish high-performance PM6/Y6/IT-4C combinations. Kranthiraja et al.\cite{14b} trained ML models on 566 polymer/NFA pairs with improved accuracy ($r = 0.85$ for RF). The polymer donors in both datasets (1,203 data points in fullerene-based OPVs and 566 in NFA-based OPVs) are fragmented into 382 distinct donor units and 526 unique acceptor units, which could potentially generate $200 \times 322$ polymers. The PCEs of OPVs constructed by screened polymers and ITIC or IT-4F acceptors are predicted by the built RF model. As shown in Figure 9a, the polymer PBDT(SBO)TzH was modified by replacing sulfur-bridged 2-butyloctyl (BO) chains (SBO) with 2-ethylhexyl (EH) on 3-chlorothiophene linked.

**Figure 6.** a) Scheme of AI design framework for developing high-performance OPV donor materials proposed by Sun et al. b) Results of ML screening: predicted PCE values obtained by GBRT versus experimental PCE values for the testing subset (170 molecules). c) Histogram of the predicted PCE values obtained by GBRT for the 18,960 new molecules in the virtual library. Reproduced with permission.\cite{31} Copyright 2021, American Chemical Society.
to BDT and varying the chains of alkylthiophene-flanked Tz (H, EH, BO, and 2-hexyldecyl (HD)) to explore the effect of alky chains on PCEs. The predicted PCE of PBDTTzEH (10.10%) is similar to its experimental value (11.15%) and the predicted PCEs of these D/A pairs (Figure 9b) are consistent with the experimental trends (EH > BO > HD > H). As shown earlier, ML-assisted HTVS under the right conditions possess advantageous predicting capabilities that speed up the development and synthesis of novel materials.

5. ML-Assisted OPV Device Optimizations

The complexity of the OPV not only comes from various molecular designs including the choice of the donor and acceptor, but also from the device technologies such as chemical synthesis, photocurrent composition, film forming, and photostability. Although in theory novel materials could be generated through ML models, it is challenging to characterize their synthesizability and experimental process. Hence, some of the latest ML approaches tackle experimental process optimization, and a brief summary (Table 3) in this trend is included in given review.

Exploring synthesizability and optimum experimental conditions for novel materials has aroused increasing interest. The computer-aided synthesis planning (CASP) program combined AI with organic chemistry to improve the likelihood of experimental success has been applied in drug or drug-like substances. The AI-driven automated synthesis for novel functional materials is attractive and challenging. ML methods are used to explore the performance maximum and stability in the experimental process to ease and speed up validation of novel chemical candidates. Data are more homogeneous (obtained...
Figure 8. a) Chemical structures of the investigated D/A pairs, including polymer donors (PM6 and PBDB-T) and NFAs (Y-ThCN, Y-ThCH3, Y-PhCl). b) J–V curves of devices based on PM6/Y-ThCN, PM6/Y-ThCH3, and PM6/Y-PhCl. c) J–V curves of devices based on PBDB-T/Y-ThCN, PBDB-T/Y-ThCH3, and PBDB-T/Y-PhCl. Reproduced under the terms and conditions of the CC BY license. [14a] Copyright 2020, The Authors. Published by Springer Nature.

Figure 9. a) Molecular structure of PBDTTz polymers for NFA-OPVs. b) Bar plots of experimental (left axis) and predicted (right axis) PCE for PBDTTz polymers. Gray bars represent PCE predicted using FP(n), FP(p), and MP(n) while the black bars represent PCE predicted using FP(n), FP(p), MP(n), and MP(p). Reproduced with permission. [14b] Copyright 2021, Wiley-VCH Verlag GmbH & Co.
typically in the same lab by modifying some experimental conditions) and denser in these automated works. While ML models have made these predictions quite accurate, it should be noted that the tuning of experimental parameters is conventional engineering practice and well grounded in the field of optimization.\(^{[76]}\) While all the examples seen in this section have only one (or very few) combination of D/A compounds, automatic experiments are much more challenging when the chemical components change. The scope of this review only focuses on several applications regarding efficiency, stability, and photocurrent. The involved fundamental definitions/concepts during high-throughput experimental processes can be found in the study by Rodríguez-Martínez et al.\(^{[15]}\)

The PCE of a given OPV system is influenced by many processing parameters. Cao et al.\(^{[77]}\) applied SVM methods with a radial basis function to fit PCE, \(J_{\text{SC}}\), and \(V_{\text{OC}}\) with total concentration, spin speed, and three different donor concentrations and increased the probability of finding a true optimum in optimizing OPV devices. The given approach can be used in searching for areas of interest of processing parameters for further optimization. Kirkey et al.\(^{[78]}\) used the same method to find optimal experimental processing parameters for all-small-molecule OPV devices consisting of one donor molecule DRCN5T and different NFAs (ITIC, IT-M, IT-4F) for each device. During their analysis, topographical maps were produced for the processing parameters related to PCE to visualize the relationships between the parameter and PCE without prior knowledge of novel materials. When the device parameters are continuous variables, the model can be represented graphically as a cross section of the multidimensional function and the results are more easily interpretable.

An et al.\(^{[77]}\) trained the RF regression model for targeting PCEs of 2,218 devices (fabricated with various compositions and thicknesses) using experimental DDs (referred to as DD, a feature that contains the thickness and ratio between components) of each material in the PM6/Y6/IT-4F ternary blends as descriptors. Predicted PCEs of all the combinations of limited DDs for PM6/Y6/IT-4F are shown in Figure 10a. The 2D graphs converted from 3D graph could clearly present different optimum compositions for different thicknesses (total DD [TDD] of \(c=20, 40, \text{ and } 60 \mu \text{g cm}^{-2}\) could be regarded as thin, middle, and thick films, respectively), which are useful for OPV devices with a specific thickness. Then, the predicted PCEs were presented with higher resolution; thus, more datasets were considered (8,000,000 datasets varying in experimental parameters without changing molecules). Datasets with PCE > 8% were filtered out and the absolute DD values were converted to relative compositions (PM6 and Y6 fraction), as shown in Figure 10b. Consequently, data points at each xy coordinate are shown in Figure 10c with the best printable formulation (BPF) at PM6/Y6/IT-4F = 1.1:1:0.17. The best efficiency formulation at PM6/Y6/IT-4F = 1.1:1:0.27 at 28.25 \(\mu \text{g cm}^{-2}\) TDD was also found. The BPF solution for experimental validation obtained a PCE of 8.85% and the BEF solution showed a higher PCE (10.2%) from roll-to-roll-produced OPVs.

The relationships uncovered by ML methods are useful not only for optimizing the device performance, but also understanding the physical roles of the parameters, which are crucial for building an interpretable QSRR model. Recently, Rodríguez-Martínez et al.\(^{[79]}\) used Bayesian and random decision forest algorithms combined with high-throughput experimental screening to rationalize photocurrent composition (referred to as \(J_{\text{SC}}\)-vol%) space for multicomponent binary NFAs-based OPV systems. The combinatorial library consisting of 24,000 data points with a large number of combinations of thickness and composition (D:A ratio) was generated by blade coating and subsequent imaging heterogeneous film features, which could be ideal seeds for ML algorithms. As demonstrated in Figure 11, 23 descriptors including optical descriptors and electronic descriptors of donor and acceptor materials, respectively, and nine descriptors built from fundamental magnitudes were selected for 8,000 data points to predict \(J_{\text{SC}}\)-vol% of eight D/A blends (each binary contains 1,000 data points in the first run) and they achieved impressive accuracy with MAE of 0.09. Despite the great prediction power of this ML method, it has some limitations including month-scale times for training, poor extrapolation, and

---

### Table 3. Comparison of several works adopting ML-assisted experiment design in OPVs.

| References | [77] | [78] | [34] | [79] | [81] |
|------------|------|------|------|------|------|
| OPV type   | FA-based bulk | NFA-based bulk | NFA-based bulk | NFA-based Binary | NFA-based Quaternary |
| Approach   | DOE  | DOE  | Automated experimentalizations | Gradients | Self-driven laboratory |
| ML algorithm | SVM  | SVM  | GPR  | BRDF\(^{[76]}\) | BNN\(^{[76]}\) |
| Casting procedure | Spin coating | Spin coating | Spin coating | Blade coating | Drop casting |
| No. data points | 235  | 41   | 100  | 24,000 | 288  |
| feature     | Processing parameter\(^{[7]}\) | Processing parameter\(^{[7]}\) | Processing parameter\(^{[7]}\) | Processing parameter\(^{[7]}\) | Processing parameter\(^{[7]}\) |
| No. ALs     | 235  | 41   | 80   | 15    | >2,000 |
| Target      | PCE  | PCE  | Photostability/PCE | \(J_{\text{SC}}\)-vol% | Photostability |
| Reproducibility | /    | /    | High | High | Very high |
| Platform    | /    | /    | AMANDA Line One | /     | ChemOS |

\(^{[7]}\)Bayesian random decision forest; \(^{[7]}\)Bayesian neural network; \(^{[7]}\) Spin speed, total concentration, donor fraction, concentration of additive; \(^{[7]}\) Solution concentration, donor fraction, temperature and annealing duration; \(^{[7]}\) D:A ratio, concentration, spin speed, solvents additives and their volume, annealing temperature and time for AL, material in the ETL, ETL annealing temperature and time; \(^{[7]}\) D:A ratio, thickness; \(^{[7]}\) Weight fraction of materials in AL.
uninformative utilization of descriptors. Thus, they adopted RF algorithm training on eight D/A blends with 23 descriptors to improve prediction capability. By extrapolating a single RF model trained on unseen D/A blends, they obtained MAE < 0.20 in
seven D/A binaries and MAE at 0.17 in 15 D/A binaries out of the training set, which means that this model may have good extrapolating performance. In addition, the bandgaps of donor and acceptor materials were found to extrapolate moderately well (MAE $\approx 0.16$) in some unseen D/A blends with RF models. Nevertheless, the proposed model fails in predicting $J_{SC}$ of PTB7-Th binary due to the absence of highly semicrystalline systems in the training set.

In fact, besides PCE, other device properties like $V_{OC}$, $FF$, $J_{SC}$, and stability can be also modeled and optimized via ML methods. David et al. used supervised learning in a sequential minimal optimization regression model (SMOreg) training on a dataset containing 1,850 entries of device properties (e.g., substrate type, environmental conditions, light type, temperature, and relative humidity) to predict stability and the initial PCE of OPV devices with $r > 0.70$. They provide methods for material identifications in terms of improved stability and top performance. Langner et al. studied photostability for four-component AL blends (P3HT/PCBM/IDTBR/PTB7-Th or PBQ-QF), based on an extended ChemOS software platform which supports Bayesian neural networks (BNNs) ML-assisted experiment planning for automated experiments. BNNs are trained on 1,041 experiments for PBQ-QF and PTB7-Th blend systems to predict the photodegradation measured for individual polymer blends (PBQ-QF and PTB7-Th blend system) by the different compositional distances of weight fractions for four components. The results produced correlation coefficients of 0.88 (0.87) on the PBQ-QF (PTB7-Th) blend system, indicating good prediction ability of the BNNs. In addition, the results suggest that PBQ-QF blends improved the stability over PTB7-Th blends.

Recently, Du et al. provided the high-throughput automated platform, AMANDA Line One, which could achieve reliable screening of photostability and efficiency for PM6/Y6 OPV materials and devices within 70 h (Figure 12a,b). The processing parameters such as D:A ratio, spin speed, annealing temperature, additives, ETL material, and other processing parameters (Figure 12c) could be automatically investigated on the platform. The part of the platform used for layer deposition (Figure 12d), layer image by the camera, UV–vis absorption, current density-versus-voltage ($J$–$V$) measurements, and offline degradation testing (Figure 12e) could be automatically achieved. The device reproducibility (Figure 12f) in terms of performance and optical properties is relatively good and consistent. As

---

Figure 11. The photocurrent–composition prediction workflow for binary OPV blends is divided into three main blocks. First, the generation of parametric libraries by blade coating on functional devices in the form of lateral gradients in the AL thickness and the D:A ratio. Second, the high-throughput photovoltaic characterization by means of colocal Raman spectroscopy and photocurrent imaging, which serves to correlate the local device performance with the variation of the target features (thickness and D:A ratio). Third, AI algorithms are trained on the experimental datasets using intrinsic fundamental descriptors of the blended materials. In the last step, the AI models are exploited to make predictions of the photocurrent–composition dependence for materials in and outside of the training dataset. Reproduced under the terms and conditions of the CC BY license. Copyright 2021, The Authors. Published by the Royal Society of Chemistry.
presented in Figure 12g, GPR analysis was performed on \( \approx 100 \) variations (including variations in D:A ratio, concentration, spin speed, annealing temperature/time for AL and ETL, additives, additive volume, and ETL material) of PM6/Y6 OPV devices with measuring optical features as descriptors for predictions of photovoltaic parameters (PCE, \( V_{OC} \), FF, and \( J_{SC} \)). The results show that the ratio of \( RMSE_{\text{train}}/RMSE_{\text{test}} \) was between 0.9 and 1.2, implying good generalization. Also, RMSE of 10 meV for \( V_{OC} \) is obtained, implying reliability of AMANDA Line One. The GPR analysis was also conducted for predicting photostability by optical features on 48 processing conditions with 84 OPV devices. They found that all devices with initial PCE > 12% and burn-in losses < 10% after 50 h from stability screening featured a high spin speed and annealing temperature of AL < 100 °C. Owing to the generality, this kind of automated research line can be potentially extended to other research fields like PV technologies, photodetectors, and transistors, enabling self-driven experimentation feasible.

6. Conclusion and Outlook

ML models as discussed in this review have been increasingly integrated into the workflow of predicting properties, screening candidates, and optimizing devices for OPV systems. In general, many kinds of ML models are trained with a large dataset (Table 1) collected by numerous efforts from the experimental literature to probe the best regression/classification prediction accuracy. Simple structural information has shown great potential in OPV performance predictions. However, the microscopic electronic structure properties such as energy levels, absorption ability, and molecular size are important in explaining photovoltaic phenomena and accordingly building the interpretable ML models. Moreover, a dataset with sufficiently large size, diversity, and homogeneity is necessary for achieving a satisfactory prediction accuracy. Many of the reported works have achieved reasonable prediction accuracies and a few virtual screening works based on ML models with experimental validations have been
conducted with promising outputs. In addition, ML methods also help the optimizations of experimental process parameters successfully. To further improve the performance and stability of new OPV devices, a plethora of works could be conducted with data-driven methods.

First, more attention should be paid to the extrapolation ability of ML models for OPV. Most of the currently proposed models can predict molecules with known units or blocks in the dataset. However, extrapolations on completely novel donor, acceptor, and D/A pairs during the predicting performance and device optimization process are highly challenging but very attractive.

Moreover, because of the small size of the experimental OPV dataset, which prevents the application of more powerful ML algorithms like deep learning, the choice of descriptors is crucial in ML-OPV studies, for building an ML model with better generalizability and interpretability. However, some computational or experimental parameters (such as solvent type, additives, crystallinity, D:A ratio, processing temperature, etc.) directly related with OPV performance are nontrivial to obtain. Thus, increasing the size of the dataset (especially improving the types of NFA structures) may improve the prediction accuracy but the data are collected by many laboratories and it is not possible to uniformly collect data on the miscibility/solubility of new materials, often available in small quantities in a single lab. For these reasons, it can be convenient to use computed properties that describe complex properties of OPV.

Furthermore, inverse design (combining ML and HTVS) of OPV materials with targeted properties both in donor and acceptor can be further explored for chemical exploration in a large space. The developing of more technical fingerprints of OPV materials could accelerate this process, increasing chances to find completely new materials with purpose-specific properties.

In addition, experimental verifications of novel D/A pairs screening by HTVS under the help of ML models are highly desirable. The challenging problem for designing novel materials is the estimation of synthesizability and choice of fabrication protocols. Several reported successful examples already indicated the tremendous potential of data-driven methods in OPV performance predictions. The recent progress of utilizing self-driven laboratories combining ML-assisted virtual screening, automated synthesis, and high-throughput characterization opens a fascinating new window for the future revolution of the material development framework.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (no. 22073045) and the Fundamental Research Funds for the Central Universities. The authors thank Dr. Yaping Wen and Dr. Luis Vasquez for helpful discussions.

Conflict of Interest

The authors declare no conflict of interest.
Lastra, K. W. Jacobsen, K. S. Thygesen, M. N. Schmidt, J. Chem. Phys. 2018, 148, 241735.

[69] a) S. P. Peng, Y. Zhao, J. Chem. Inf. Model 2019, 59, 4993; b) G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, arXiv preprint arXiv:1705.10843 2017.

[70] Y.-C. Lin, Y.-J. Lu, C.-S. Tsao, A. Saeki, J.-X. Li, C.-H. Chen, H.-C. Wang, H.-C. Chen, D. Meng, K.-H. Wu, Y. Yang, K.-H. Wei, J. Mater. Chem. A 2019, 7, 3072.

[71] T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan, J. Kettle, Nano Energy 2020, 78, 105342.

[72] F. Hase, L. M. Roch, P. Friederich, A. Aspuru-Guzik, Nat. Commun. 2020, 11, 4587.

[73] C. W. Coley, W. H. Green, K. F. Jensen, Acc. Chem. Res. 2018, 51, 1281.

[74] C. W. Coley, D. A. Thomas 3rd, J. A. M. Lummis, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotto, W. H. Green, A. J. Hart, T. F. Jamison, K. F. Jensen, Science 2019, 365, eaax1566.

[75] M. A. Soldatov, V. V. Butova, D. Pashkov, M. A. Butakova, P. V. Medvedev, A. V. Chernov, A. V. Soldatov, Nanomaterials 2021, 11, 619.

[76] J. Martins, A. Ning, Engineering Design Optimization, Cambridge University Press, Cambridge, UK 2021.

[77] B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Lubber, B. C. Olsen, A. Mar, J. M. Buriak, ACS Nano 2018, 12, 7434.

[78] A. Kirkey, E. J. Lubber, B. Cao, B. C. Olsen, J. M. Buriak, ACS Appl. Mater. Interfaces 2020, 12, 54596.

[79] X. Rodríguez-Martínez, E. Pascual-San-José, Z. Fei, M. Heeney, R. Guimerà, M. Campoy-Quiles, Energy Environ. Sci. 2021, 14, 986.

[80] Q. Burlingame, M. Ball, Y.-L. Loo, Nat. Energy 2020, 5, 947.

[81] S. Langner, F. Hase, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, C. J. Brabec, Adv. Mater. 2020, 32, e1907801.

[82] L. M. Roch, F. Hase, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, PLoS One 2020, 15, e0229862.

[83] M.-H. Lee, Energy Technol. 2019, 8, 1900974.

Zhi-Wen Zhao is a lecturer at Hubei University of Art and Science. She was a research associate in the Department of Chemistry at Nanjing University. She received her Ph.D. from Northeast Normal University in 2021. She joined as a visiting Ph.D. student in the group of Alessandro Troisi at University of Liverpool supported by the China Scholarship Council scholarship (2019–2020). Currently, her research interest lies in material discovery by data-driven methods for optoelectronic materials, targeting their structural and physical properties.

Yun Geng is an associate professor at Northeast Normal University. She received her Ph.D. from Northeast Normal University in 2011. She was a visiting scholar at the University of Warwick in 2016–2017. Her research interests broadly fall into theoretical investigations on the electron processes in organic optoelectronic materials including the organic solar cell materials, thermally activated delayed fluorescence (TADF) materials, and organic phosphorescence materials, along with HTVS for the above functional materials.

Alessandro Troisi obtained his Ph.D. in Bologna, performed postdoctoral research at Northwestern University, and started at the University of Warwick in 2005 as assistant professor, being promoted in 2010 to a full professorship. He joined the Materials Innovation Factory of the University of Liverpool in 2017. He specializes in theoretical and computational chemistry, computer-aided materials discovery, electron transport and transfer, organic semiconductors, and quantum dynamics. He was awarded the Marlow Medal by the Royal Society of Chemistry in 2007 and the ERC Starting Grant (2009), ERC Consolidator Grant (2014), and ERC Advanced Grant (2021).

Haibo Ma is professor of theoretical chemistry at Nanjing University. He has a B.S. and Ph.D. from Nanjing University in 2002 and 2007, respectively. He has received the Humboldt research fellowship from the Alexander von Humboldt foundation (2007–2009) and Tang Au-Qing youth award in theoretical chemistry from Chinese Chemical Society (2018). His main research interests focus on the development of new quantum chemical methods for strongly correlated systems and the theoretical investigation of excited-state processes in optoelectronic materials as well as molecular design via machine learning and high-throughput virtual screening for functional materials.