Solar RRL
www.solar-rrl.com

Machine Learning Approaches for Predicting Power Conversion Efficiency in Organic Solar Cells: A Comprehensive Review

Yang Jiang, Chuang Yao, Yezi Yang, and Jinshan Wang**

Organic solar cells (OSCs), renowned for their lightweight, cost efficiency, and adaptability nature, stand out as a promising option for developing renewable energy. Improving the power conversion efficiency (PCE) of OSCs is essential, and researchers are delving into novel materials to achieve this. Traditional approaches are often laborious and costly, highlighting the need for predictive modeling. Machine learning (ML), especially via quantitative structure–property relationship (QSPR) models, is streamlining material development, with a goal to exceed a 20% PCE. In this review, the application of ML in OSCs is explored, and recent studies utilizing ML approaches for PCE prediction are reviewed, encompassing empirical functions, ML algorithms, self-devised ML frameworks, and the combination with automated experimental technologies. First, the benefits of ML in predicting PCE for OSCs are addressed. Second, the development of high-efficiency predictive models for both fullerene and nonfullerene acceptors is delved into. The impact of various ML algorithm models on PCE prediction is then assessed, taking into account the construction of predictive models. Moreover, the quality of databases and the selection of descriptors are considered. Databases and descriptors based on experimental studies are further categorized. Finally, prospects for the future development of OSCs are proposed.

1. Introduction

Currently, organic solar cells (OSCs) show considerable promise across a variety of applications. These cells are advantageous as ecological energy conversion devices, offering benefits such as lightweight design, high flexibility, and cost efficiency.[1–4] However, OSCs still trail behind silicon solar cells in terms of power conversion efficiency (PCE).[5–7] Therefore, the main focus of research in this field is to enhance PCE. The efficiency of OSCs is influenced by several factors, including material properties, environmental conditions, and device structure, making it crucial to predict and optimize these factors for improved performance.[8–10] The challenge lies in deeply understanding the complex relationship between the structural configuration and functional attributes of OSCs to identify innovative and effective solutions. The optoelectronic materials for OSCs have a broad range and diverse properties.[11] Discovering or designing novel high-efficiency OSCs requires extensive experimentation with numerous compounds, followed by thorough performance analysis and rigorous feasibility verification. Traditional experimental research primarily involves designing and synthesizing new donor/acceptor (D/A) materials or optimizing preparation conditions to find the optimal operating mechanism.[12] In recent years, the study of electron acceptors in the active layer, including fullerene acceptors (FAs) and nonfullerene acceptors (NFAs), has attracted significant attention. FAs and their derivatives have a limited and weak absorption spectrum for sunlight and offer limited modifiability of molecular orbital energy levels.[13–15] As a result, the PCE of FAs OSCs is often not optimal. In contrast, NFAs offer notable advantages such as adaptable synthesis, tunability of absorption and electron energy levels, and strong absorption of visible light and near-infrared solar radiation. Research on NFAs is advancing towards achieving a PCE of 20%.[16–19]

PCE is achieved through manual design, material engineering, and device engineering (**Figure 1**). The development of OSCs involves an empirical screening process, and the traditional "trial and error" approach to discovering new materials is complex and time consuming, potentially leading to resource wastage and high costs. In the study of materials' physical and chemical properties, density functional theory (DFT) calculations are performed based on the materials' fundamental principles.[20] However, most computational methods are specific to certain systems and are not well suited for material systems with complex and varied crystal structures, showing significant limitations. With the advancement of big data and artificial intelligence (AI), machine learning (ML), which is centered around data-driven algorithms, can aid research in the field of OSCs.

Y. Jiang, C. Yao, Y. Yang
Chongqing Key Laboratory of Extraordinary Bond Engineering and Advance Materials Technology (EBEAM), School of Materials Science and Engineering
Yangtze Normal University
Chongqing 408100, P. R. China
E-mail: yaochuang@yznu.cn

Y. Jiang, J. Wang
School of Materials Science and Engineering
Yancheng Institute of Technology
Yancheng 224051, Jiangsu, China
E-mail: wangjinshan@ycit.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/solr.202400567.

DOI: 10.1002/solr.202400567

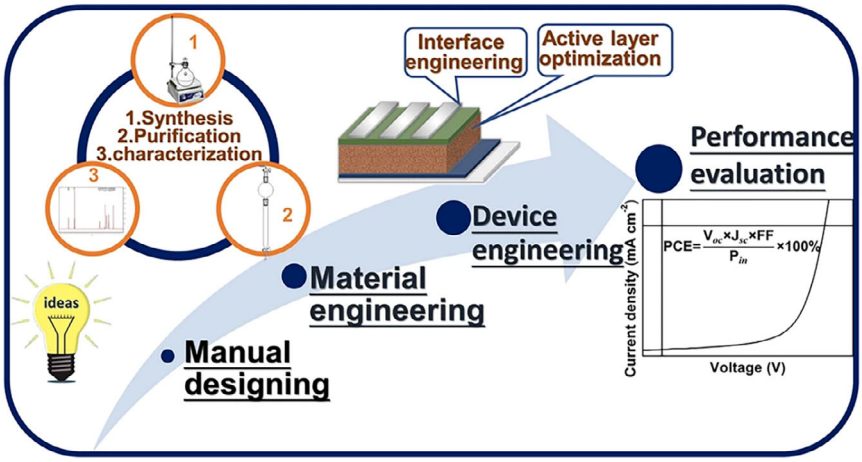

Figure 1. The process of traditional experiments in the study of OSCs. Reproduced with permission.[58] Copyright 2020, Springer Nature.

By employing the appropriate algorithmic model, material performance can be quickly predicted, thus speeding up the screening and design processes for new materials (**Figure 2**a). To enable high-throughput screening and rational design of efficient organic photovoltaic (OPV) materials, identifying the most relevant molecular descriptors for photovoltaic performance is essential.[21] This task is a formidable challenge because OSCs depend on a variety of optoelectronic processes. For instance, it involves quantifying the correlation between molecular descriptors and the PCE, as well as with the fundamental photovoltaic parameters, namely, open-circuit voltage (V_{OC}), short-circuit current density (J_{SC}), and fill factor (FF).

This innovative approach, integrating ML, stands in contrast to the traditional "trial and error" methodology, which can take hundreds of hours to evaluate a compound through experimental procedures. This new technology significantly enhances the screening efficiency of OSCs through the use of corresponding algorithms. It accelerates the development of new materials, thereby substantially reducing the time and costs associated with material design. A comprehensive statistical analysis of publications from 2014 to 2023, conducted using the Web of Science search engine and the keywords shown in Figure 2b, reveals that ML has become a rapidly growing field with an increasing number of applications in material research. The exponential growth in the number of publications with the combined keyword "machine learning + organic solar cells" over the past decade underscores this trend. It is anticipated that ML will have extensive application prospects in the field of OSCs, offering vital support and assistance in new material design, experimental optimization, data analysis, simulation prediction, and other areas.

In this review, we briefly discuss the successful application of ML approaches in predicting properties of OSCs devices in recent years. In Section 2, we mainly introduce the modeling predictive methods of different ML technologies for OSC devices, including the use of empirical functions, ML algorithms, and self-devised ML frameworks, as well as the combination of automated experimental technology with ML. Then, various ML algorithms were discussed, revealing a common phenomenon: Random Forest (RF) models typically achieve the best performance in predicting OSCs. In Section 3, we emphasize the importance of building suitable databases. Section 4 categorizes ML input descriptors into molecular, fingerprint, microscopic property, and image-based types. Section 5 discusses the

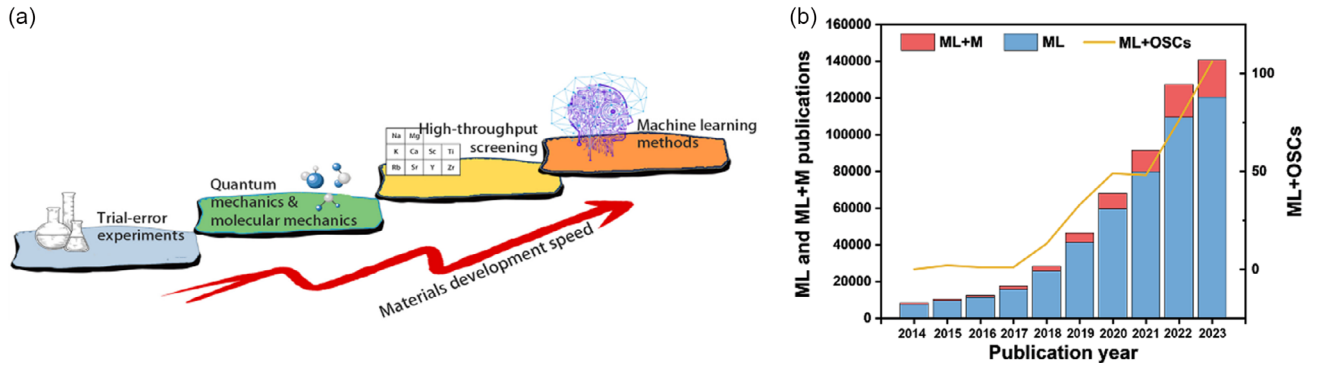

Figure 2. a) Development in methods to accelerate new materials discovery. Reproduced with permission.[154] Copyright 2020, John Wiley and Sons. b) The number of published works pertaining to "machine learning (ML)," "machine learning + material (ML + M)," and "machine learning + organic solar cells (ML + OSCs)" over the past decade.

inherent limitations of ML methods in predicting the efficiency of OSCs, including challenges in model construction, database quality, and descriptor selection. Addressing these challenges has enhanced our understanding of the current research landscape, and we aim to offer potential research directions for photovoltaic professionals.

2. Machine Learning Predictive Methods

2.1. Machine Learning Approaches and Evaluation Metrics

ML approaches enable computer systems to learn from data, make decisions, and predict outcomes. ML is primarily categorized into regression, deep learning, generative, and other types, each with unique application scenarios and advantages.

Regression analysis is a fundamental technique for predicting continuous numerical values. It forecasts future data points by modeling the relationship between input and output variables. Linear regression (LR) is a straightforward method, well-suited for capturing linear relationships. However, when data displays nonlinear characteristics, more sophisticated models such as polynomial regression or decision trees (DT) may be required. Deep learning enables the learning of complex patterns through multilayered neural networks. It performs exceptionally well in domains such as image and speech recognition, natural language processing, and more. Key deep learning architectures include deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN). These networks can automatically extract features and learn, but they typically demand substantial data and computational resources. Generative models aim to learn data distributions and generate new data instances. Generative adversarial networks (GAN) produce data by engaging in adversarial interactions between generators and discriminators. Variational autoencoders (VAE), conversely, generate new data points by optimizing the latent space distribution. These methods have broad applications in areas such as image generation, style transfer, and data augmentation.

In terms of optimization methods, gradient descent is the most common algorithm, which updates parameters by calculating the gradient of the loss function relative to the model parameters to minimize the loss. There are various variants of gradient descent, including stochastic gradient descent (SGD), mini-batch gradient descent, and momentum-based gradient descent, which enhance training efficiency and model performance through different update strategies. Nongradient methods, such as genetic algorithms (GA), simulated annealing, and particle swarm optimization, do not rely on gradient information. They are suitable for situations where gradients are difficult to compute or the problem space is nonconvex. These methods explore the parameter space and search for the global optimal solution by simulating natural processes. Although highly effective in certain situations, they typically require more computing resources and time.

Meanwhile, during the ML process, once the model training halts upon reaching a predetermined number of iterations or level of accuracy, it is crucial to assess the performance of the predictive model across various parameters.[22] For example, larger datasets often necessitate more iterations for complete learning, and complex models may require additional iterations for optimization. Meanwhile, the available computing resources dictate the number of iterations feasible. To prevent overfitting, researchers can halt training when the validation set's performance plateaus.[23] For supervised learning, the initial step involves establishing the model evaluation metrics.[24] The evaluation indicators for regression problems include mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). The calculations are shown in Equation (1)–(4).

$$\mathrm{MAE} = \frac{1}{n}\sum |y_{\mathrm{real}} - y_{\mathrm{pred}}| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (y_{\mathrm{real}} - y_{\mathrm{pred}})^2} \tag{2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum \left|\frac{y_{\mathrm{real}} - y_{\mathrm{pred}}}{y_{\mathrm{real}}}\right| \cdot 100\% \tag{3}$$

$$R^2 = 1 - \frac{\sum (y_{\mathrm{real}} - y_{\mathrm{pred}})^2}{\sum (y_{\mathrm{real}} - \bar{y})^2} \tag{4}$$

In the formula, y_{real} is the actual value of photoelectric conversion efficiency, %; y_{pred} is the predicted value of photoelectric conversion efficiency, %; $\bar{y}$ is the average value of actual photoelectric conversion efficiency, %; and n is the number of samples.

RMSE is a widely used metric for evaluating the precision of regression model forecasts. It is commonly employed as a loss function to optimize model parameters and enhance predictive accuracy by minimizing RMSE. RMSE also offers an intuitive measure of error, aiding in understanding the magnitude of errors. Conversely, MAE quantifies the average magnitude of prediction errors, serving as a clear indicator of model performance. Both metrics provide intuitive insights into the magnitude of a model's prediction errors. Since they share the same units as the y variable, they are well-suited for assessing the model's prediction accuracy in specific units. As R^2 is dimensionless, it is ideal for comparing models, particularly across different datasets or units, because it offers a standardized measure of performance.

The smaller the MAE and RMSE, the higher the accuracy of the predictive model. The range of MAPE is from 0 to positive infinity; the closer it is to 0, the higher the accuracy of the prediction. When the predicted value is exactly the same as the actual value, the MAPE is 0. When the difference between the predicted value and the actual value is larger, the MAPE is larger. The closer R^2 is to 1, the closer its prediction results are to the true value, indicating that the fitting effect of the selected model is better.

Pearson correlation coefficient r is also commonly used as a regression evaluation index between predicted values and true values in OSCs. Similar to R^2, the closer r is to 1, the more accurate the model prediction is. The formula is as follows:

$$r = \frac{\mathrm{Cov}(y_{\mathrm{real}}, y_{\mathrm{pred}})}{\sqrt{\mathrm{Var}(y_{\mathrm{real}})\mathrm{Var}(y_{\mathrm{real}})}} \tag{5}$$

Here, y_{real} and y_{pred} represent the actual and predicted values, respectively, and $\mathrm{Cov}(y_{\mathrm{real}}, y_{\mathrm{pred}})$ is the covariance between the

actual and predicted values. A dimensionless correlation coefficient, ranging from −1 to 1, can be derived. This coefficient reflects the degree of linear correlation between the two variables.

Dividing the available data into distinct training and test sets allows for the estimation of a model's performance. Repeatedly selecting different subsets of observations enables multiple assessments of the model's performance, with the outcomes from each training and testing phase being aggregated.[25] This approach also serves to measure the variability and stability of the model's efficacy. To execute this strategy, cross-validation (CV) and guidance programs can be employed.[26] The general workflow for training models in ML is depicted in **Figure 3**. In materials science and discovery, this workflow encompasses several stages: preparing one or more material databases, choosing suitable input features, constructing precise prediction models, forecasting and assessing target attributes, and ultimately validating and refining ML models.[27]

2.2. Empirical Function Predictive Models

Based on the classical thin-film solar cell concept, a metal–insulator–metal (MIM) model was employed to investigate the bulk heterojunction (BHJ) devices.[28] It was discovered that polymer-fullerene solar cells are influenced by the morphology of the active layer.[29] However, even the extended MIM model is not fully adequate to explain the V_{OC} of BHJ solar cells.

Gadisa et al.[30] examined the electrochemical characteristics of six similar polythiophene derivatives in relation to the V_{OC}, using PCBM as the electron acceptor and the same polythiophene derivatives as the electron donors. Their goal was to establish a correlation between these properties. Scharber et al.[31] developed a computational model that correlates the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energy levels of polymers with the maximum V_{OC} efficiency in BHJ solar cells. This model, based on empirical data, provides valuable insights into the optimal material parameters for conjugated polymer and PCBM devices. The solar spectrum integral is calculated from 0 to the optical bandgap of the donor material, measured in eV. The FF is held constant at 65%. Derived from an extensive study of 26 BHJ devices with PCBM exclusively as the acceptor, this model assumes that all photons with energy above the bandgap are absorbed, ignoring the acceptor's photon absorption and the oscillator strengths of the absorbers. Alharbi et al.[32] proposed a predictive function model to effectively screen potential solar cell materials, incorporating additional factors such as the acceptor's optical bandgap and exciton diffusion length, which enhances the accuracy of efficiency predictions. However, the calculation of the absorption spectrum remains unclear in the model's reconstruction. When calculating the PCE using the Alharbi model, the J_{SC} equation from the Scharber model is applied. It should be emphasized that these function models primarily cater to FA solar cells.

In early studies of OSCs, the focus was mainly on FAs and their derivatives. Currently, research in OSCs has shifted towards polymer donors and small molecule acceptors.[33–35] The advent of NFAs has demonstrated significant potential for advancing OSCs, offering improved performance and reduced costs.[36–41] NFAs are divided into two main groups: Rylene-diimides and fused ring electron acceptors (FREAs). The ongoing refinement of FREAs has addressed critical issues that traditional FAs could not.[42] Furthermore, the introduction of nonfused ring electron acceptors (NFREAs), with their simpler molecular structures and easier synthesis, has significantly propelled the OSCs field forward.[43–45] This has led to a new era dominated by NFA technologies.[46–48] To improve the prediction of NFAs' material efficiency, Imamura et al.[49] refined the Scharber model. They theoretically explored new NFA materials, modeled the interfacial geometry between NFAs and polymers, and assumed FF

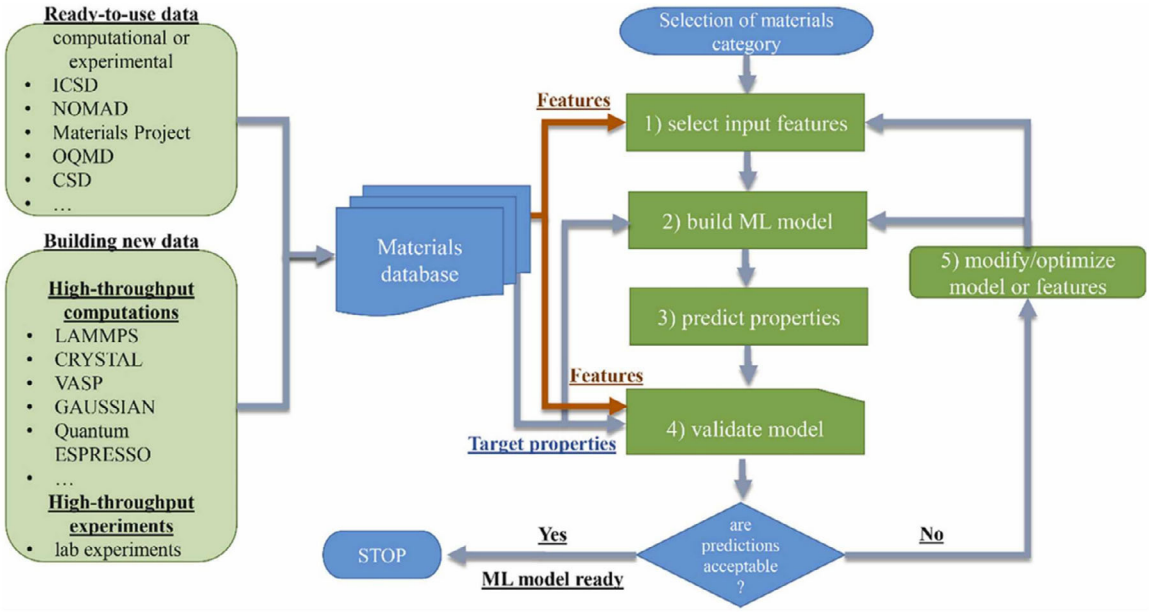

Figure 3. Schematic diagram of training an ML model in materials science. Reproduced with permission.[27] Copyright 2023, Elsevier.

of 70%. Due to insufficient photon absorption or low charge separation and transmission efficiency, J_{SC} may decrease. Therefore, J_{SC} calculations take into account possible energy losses by including a term for the energy difference between donor LUMO and acceptor LUMO. While this model offered guidance for the development of polymer donors and NFAs, it was not an effective predictor for high-performance OPV devices based on NFAs at the time. The PCE of the model can be calculated using the formula $\mathrm{PCE} = V_{OC} \times J_{SC} \times \mathrm{FF}$.

Hutchison et al.[50] proposed a more precise model for estimating the efficiency of NFAs-OSCs, known as the OPEP model, which includes OPEP/B3LYP and OPEP/sTD-DFT. The OPEP/B3LYP model employs time-dependent density functional theory (TD-DFT) to calculate the excitation energies and oscillator strengths of electronic excited states. In contrast, the OPEP/sTD-DFT uses a model based on semiempirical simplified TD-DFT, which accelerates the computation process and efficiently predicts PCE. However, the computation time for both models depends on several factors, including molecular system size, the accuracy of the basis group chosen, computational resources (such as central processing unit core count and memory size), and the extent of program optimization. The mathematical formulations for these typical functional predictive models are detailed in **Table 1**.

The equation to predict PCE by using B3LYP (OPEP/B3LYP) and sTD-DFT (OPEP/sTD-DFT) for high experimental PCE (PCE experimental values above 9%)

$$\begin{aligned}\mathrm{PCE} = {} & -136.5 + (6.42\,\mathrm{eV}^{-1})E_{T_1} + (0.917\,\mathrm{D}^{-1})\Delta\mu_{ge} \\ & + (8.31\,\mathrm{eV}^{-1})\omega_D^- - (0.00389\,\mathrm{cm})E^D_{\mathrm{OSCS}-10} \\ & - (0.00047\,\mathrm{cm})E^A_{\mathrm{OSCS}} - (0.0036\,\mathrm{au}^{-1})\alpha^A\end{aligned} \tag{6}$$

$$\begin{aligned}\mathrm{PCE} = {} & 99.83 - (7.53\,\mathrm{eV}^{-1})\Delta E^A_L + (5.49\,\mathrm{eV}^{-1})\Delta E^A_H \\ & - (1.66\,\mathrm{D}^{-1})\mu^D_g + (1.57\,\mathrm{arb\,unit}^{-1})\sum f^A \\ & + (0.098\mathrm{arb\,unit}^{-1})\,Abs^D_{\mathrm{FOM}} - (0.0047\,\mathrm{cm})E^D_{\mathrm{OSCS}-10}\end{aligned} \tag{7}$$

In Table 1 and Equation (6) and (7), $E^{\mathrm{Donor}}_{\mathrm{HOMO}}$ is the HOMO energy level of the donor; $E^{\mathrm{PCBM}}_{\mathrm{LUMO}}$ is the energy level between the LUMO of the PCBM; 0.3 V is the bandgap offset; e and q are the elementary charge; $\Delta E^{\mathrm{donor}}_{\mathrm{gap}}$ is the bandgap energy of the donor; $\varnothing_{\mathrm{ph}}(\lambda)$ is the photon flux density as a function of the wavelength λ of light; ΔE_{LUMO} represents the LUMO energy of the acceptor; 0.85, 0.1, and 0.03 are empirical parameters used in the model to adjust the calculation of J_{SC}; E_g represents that the bandgap energy directly affects V_{OC} of the device; 0.5 represents the voltage loss caused by nonideal factors within the device, such as charge recombination and energy level mismatch; $0.0114E_g^{1.8617}$ and $0.057E_g$ are empirical formulas based on experimental data, used to further adjust the calculation of V_{OC}; $\varnothing_{\mathrm{ph}}(E)$ represents the photon flux density at energy E, which is usually related to the distribution of the solar spectrum; E_{gap} represents the electrons excited by photons in the bandgap that contribute to J_{SC}; $\alpha(E)$ represents the material's ability to absorb light, which determines the probability of the material absorbing photons; L_d is the thickness of the charge transport layer in the device, which affects the separation and collection efficiency of photogenerated charges; k_B is the Boltzmann constant; and T is temperature.

The following are specifically designed for calculating OPEP/B3LYP and OPEP/ sTD-DFT models: $E^D_{\mathrm{OSCS}-10}$ and $E^A_{\mathrm{OSCS}-10}$ are the D/A materials with the highest oscillator strength that can transition from the ground state to the excited state in the first 10 transitions, μ^D_g and μ^A_g are the net dipole moment calculated from the ground state dipole moment vectors from the TD-DFT or sTD-DFT calculations, ΔE^D_H and ΔE^A_H are the energy difference in eV between the HOMO and HOMO − 1 levels, ΔE^D_L and ΔE^A_L are the energy difference in eV between the LUMO and LUMO + 1 levels, ω^- and ω^+ are the electrocoating power is the tendency of the molecule to D/A electrons, ω is the electrophilicity index used to describe the change in energy when a molecule suddenly enters an electron-rich environment transfer, and $\Delta\mu_{ge}$ is the change in dipole moment from going from the ground state to the first excited state. The excited state dipole moment was calculated from an energy gradient calculation with TD-DFT, f_0 is the oscillator strength of the first transition from the ground state to the first excited state, E_{T_1} is the electronic transition energy from the singlet ground state to the first triplet excited state and it was only calculated for the OPEP/B3LYP model, $\sum f^A$ is the sum of oscillator strengths in the spectrum, N^A_{atom} is the conjugation length of the molecule is calculated by counting the number of atoms in the conjugation path, α is the

Table 1. The mathematical formulations for the typical function predictive models.

	V_{OC} [V]	J_{SC} [$\mathrm{mA^{-1}\ cm^2}$]	FF [%]
Scharber	$\frac{1}{e}(\lvert E^{\mathrm{Donor}}_{\mathrm{HOMO}}\rvert - \lvert E^{\mathrm{PCBM}}_{\mathrm{LUMO}}\rvert) - 0.3\,\mathrm{V}$	$0.65\int_0^{\Delta E^{\mathrm{donor}}_{\mathrm{gap}}} \varnothing_{\mathrm{ph}}(\lambda)d\lambda$	65
Imamura	$\frac{1}{e}(\lvert E^{\mathrm{Donor}}_{\mathrm{HOMO}}\rvert - \lvert E^{\mathrm{PCBM}}_{\mathrm{LUMO}}\rvert) - 0.3\,\mathrm{V}$	$\frac{0.85}{\exp\left[\frac{-(\Delta E_{\mathrm{LUMO}}-0.1}{0.03}\right]+1}\int_0^{\Delta E^{\mathrm{donor}}_{\mathrm{gap}}} \varnothing_{\mathrm{ph}}(\lambda)d\lambda$	70
Alharbi	$E_g - 0.5 - 0.0114E_g^{1.8617} - 0.057E_g$	$q\int_{E_{\mathrm{gap}}}^{\infty} \varnothing_{\mathrm{ph}}(E)\left[1 - e^{\frac{-\alpha(E)L_d}{\cos\frac{\pi}{4}}}\right]$	$\frac{V_{OC}}{V_{OC}+12k_BT}$
OPEP/B3LYP (PCE > 9%)	$0.78 - 0.0052\mu^A_g - 0.021f^A - 0.029\omega_A + 0.00017\alpha^A + 2.56\times10^{-5}E^A_{\mathrm{OSCS}-10}$	$J_{SC} - 107.8 + 1.91\Delta\mu^A_{ge} - 0.0053\alpha^A + 7.85\omega_D^- + 1.23f_0 - 0.00084E_{\mathrm{OSCS}} - 10^A + 0.0036E^D_{\mathrm{OSCS}-10}$	$-178.53 + 31.86E^A_{T_1} + 1.39\omega_A^+ - 0.0083\alpha^A + 12.68\omega_D^- - 0.0017E^A_{\mathrm{OSCS}-10} + 0.0058E^D_{\mathrm{OSCS}-10}$
OPEP/sTD-DFT (PCE > 9%)	$1.68 - 2.36\times10^{-5}U^D - 0.014\mu^A_g - 0.00049N^A_{\mathrm{atom}} - 0.0013Abs^A_{\mathrm{FOM}} - 0.0054Abs^D_{\mathrm{FOM}} - 0.11\omega_{\bar{A}}$	$-22.75 - 12.89\Delta E^A_L + 7.66\Delta E^A_H + 0.00064U^D + 5.43\omega_A + 2.4\sum f^A + 0.16Abs^D_{\mathrm{FOM}}$	$184.39 - 2.55\mu^D_g + 1.53\sum f^A + 0.23Abs^D_{\mathrm{FOM}} + 15.81\eta_A - 0.008E^D_{\mathrm{OSCS}-10} - 0.0014E^A_g$

polarizability of the acceptor in units of a.u., U is the single point energy given by sTD-DFT in units of Hartree, η is the chemical hardness which describes the resistance to intramolecular charge, Abs^{D}_{FOM} is a new descriptor that is the most important in these two models, and it takes into account the oscillator strengths of the donor and the solar spectrum. The constants in the above formula are empirical values.

Figure 4 illustrates the predicted performance of high-performance OPV devices, using PCE data as a case study. This analysis examines and compares the error distribution ranges of previous predictive models, as well as the magnitudes of MAE and RMSE. The study confirms that the OPEP model can efficiently and precisely identify new, efficient D-NFA pairs, offering valuable guidance for future predictions of high-performance OSC materials.

2.3. Machine Learning Algorithm Predictive Models

Research is increasingly focused on using theoretical computations to deduce the performance characteristics of OSCs directly from their molecular structures. As computational chemistry progresses, quantum chemical methodologies, particularly DFT, have become essential tools for clarifying a material's microscopic properties based on its structural configuration. Despite these advancements, accurately predicting the PCE of OSC devices is still a complex challenge due to the complex nature of the active layer's structural composition and the complex physical and chemical dynamics involved in the photoelectric transduction processes of OSCs. Consequently, attaining precise computational predictions of OSCs performance remains a significant challenge for the scientific community. Here, we will explain and review some examples of using ML algorithms to study OSCs.

Material properties can be predicted by constructing a descriptor model using established input characteristics with ML algorithms. This approach effectively overcomes the inherent limitations of traditional linear and nonlinear methods. In 2011, Aspuru-Guzik et al.[51] developed a framework called the quantitative structure–property relationship (QSPR) model, which utilizes a straightforward LR algorithm. Using a comprehensive set of 33 molecular descriptors from donor materials as input variables, they accurately forecasted the PCE of the corresponding devices, obtaining an impressive coefficient of determination ($R^2 = 0.89$). The combination of QSPR with chemical informatics techniques may provide critical insights that enhance the molecular design of OSCs. This represents a pioneering effort in applying ML methodologies to bolster OSCs research.

In 2017, Perea et al.[52] conducted a comprehensive investigation into the phase evolution and mixing thermodynamics of polymer-fullerene composites. They used the Flory-Huggins interaction parameters to describe intermolecular interactions in polymer-fullerene composites and integrated these parameters into the training and prediction of artificial neural network (ANN) models. The ANN models were pivotal in predicting solubility parameters by analyzing surface charge distributions. These parameters contributed to the development of a significant metric that illuminated the blend stability within the polymer-fullerene systems.

In 2018, Ma et al.[53] predicted the PCE of small molecule OSCs, they applied five ML models. These models used 13 microscopic properties as input features. Notably, they found that using GA achieved $r = 0.79$. Saeki et al.[54] applied ANN and RF models to identify suitable conjugated molecules for polymer-fullerene OSCs. They utilized nearly 1000 experimental parameters, including HOMO energy levels, average molecular weights, PCE, and chemical structures sourced from literature. Their approach integrates ML with manual screening and experimental analysis, offering fresh insights into the development of new polymers. It is noteworthy that the majority of these ML techniques are applied within the realm of fullerene-based OSCs.[55]

With the ongoing advancement of NFAs, researchers are increasingly employing ML algorithms to address more extensive and complex OSCs.[56] Zhao et al.[57] developed an extensive causal CNN model in 2019, which they then adapted into a string-based molecular generative model. This specific ML

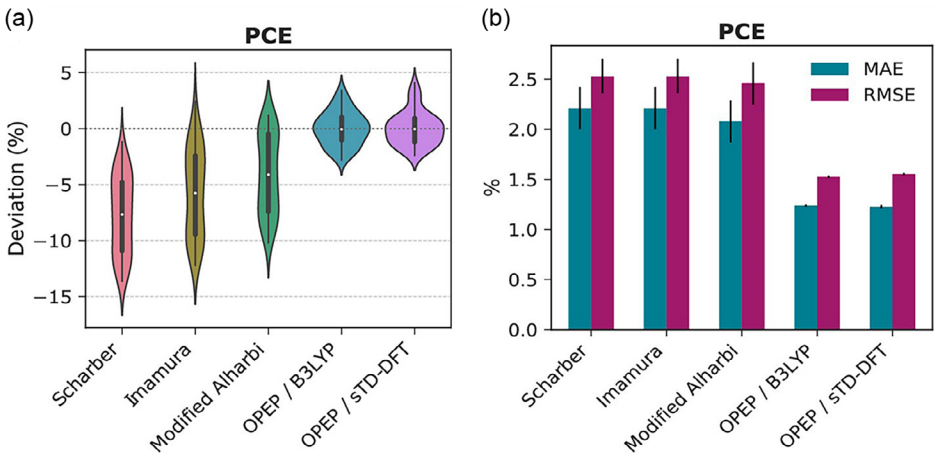

Figure 4. Comparative analysis of different prediction models for high-performance OPV devices (PCE experimental values above 9%): a) violin plots showing the distribution of deviation for PCE; b) statistical analysis of models for PCE (the error bars are from the standard error of 5-fold CV). Reproduced with permission.[50] Copyright 2022, American Chemical Society.

model serves as both a generative and predictive tool for NFAs in OSCs. The diversity of the generated NFAs is influenced by the depth of the convolutional layers. In the attribute prediction model, an enhanced CNN extracts features from the string representations of NFAs, and the predictions are subsequently verified through quantum chemical computations.

Establishing a specific ML model that links chemical structure, donor–acceptor matching systems, and photovoltaic properties allows ML model to rapidly assess and screen new materials, as well as active layer D/A pairs. For instance, in 2020, Min et al.[58] employed a comprehensive dataset comprising 565 polymer D-NFA pairs, meticulously selected from literature. A significant subset, comprising 477 D/A pairs, was dedicated to training the models. Among the ML models tested, the boosted regression trees (BRT) model proved to be the most accurate, with a notable correlation coefficient $r = 0.71$. The RF model also showed high accuracy, achieving a correlation coefficient $r = 0.7$, thereby validating the effectiveness of ensemble learning for predicting polymer D/A interactions.

Sharma et al.[59] in 2021 meticulously compiled a dataset of 154 polymers, specifically NFAs-OSCs, aimed at forecasting non-radiative voltage losses. In an effort to project the PCE of P3HT donor-based OSCs, this study employed four distinct ML models. It was determined that the RF model yielded an impressive $r = 0.857$.

In 2022, Wang et al.[60] developed a dataset of 265 NFAs using the PTB7-Th donor to predict PCE and absorption maxima. They utilized several ML models, including RF, k-nearest neighbors (KNN), logistic regression, and support vector machine (SVM). Among these, the RF model showed the best performance in predicting PCE on the validation set ($r = 0.93$). Lu et al.[61] constructed a comprehensive dataset containing NFA-related chemical structures and photovoltaic parameters obtained from literature, which served as a basis for a comparative analysis of four ML models: eXtreme Gradient Boosting (XGBoost), DT, KNN, and RF. XGBoost outperformed the others ($r = 0.79$). To gain deeper insights into the relationship between principal features and PCE, the SHapley Additive exPlanations (SHAP) method was applied. This method helped clarify the physical significance of the most influential features, providing a more nuanced and detailed understanding of the complex relationship between these key features and PCE.

Karak et al.[62] in 2023 constructed a dataset of 1242 experimentally validated D/A combinations. They used this dataset to generate material descriptors, which were then used to train and evaluate five different supervised ML models. After thorough analysis, it was found that the RF ML model had the best predictive performance for PCE ($r = 0.791$ and MAPE = 2.004%) of NFAs.

Zhang et al.[63] have established a complex relationship between the structures, properties, and photovoltaic performance of OSCs, significantly improving the PCE of NFAs. They utilized four ML models and found that the RF model had the highest accuracy and stability in predicting PCE. Additionally, they conducted a virtual screening of the designed acceptors based on the predicted PCE values. Afterwards, in 2024, Zhang et al.[64] employed four ML models: RF, extra tree regression (ETR), gradient boosting regression tree (GBRT), and adaptive boosting. The RF model demonstrated excellent generalization capabilities. The photovoltaic performance of the recombined D: A1: A2 ternary OSCs was predicted and screened, yielding 138 sets with a PCE exceeding 18.50%. The top-performing predicted PCE was 18.83%, corresponding to the D18: AQx-18: ZH2 ternary OSCs. These findings offer valuable insights for the optimization of ternary OSCs, potentially accelerating the development of high-performance devices.

Recently, Janjua et al.[65] have recently compiled a dataset of 700 OPV through meticulous data collection from various literature sources. The integration of ML techniques, cluster analysis, and synthetic accessibility has led to a highly productive and organized framework for designing polymers specifically for OSCs. Chen et al.[66] constructed a database of 310 pairs of donors and NFAs, from which they selected 39 molecular structure descriptors. They applied four ML algorithms: RF, ETR, GBRT, and Adaptive Boosting, to forecast photovoltaic parameters. The RF model demonstrated outstanding predictive accuracy. With the trained RF model, it was predicted that the PCE of 42 D-NFAs pairs surpasses 16%. Their research effectively bridged the gap between molecular structure, material properties, and photovoltaic performance of D/A materials, leading to enhanced PCE in OSCs.

Upon analyzing the provided examples, an intriguing pattern emerges: irrespective of data volume and feature inputs, ML models used to forecast the efficiency of OSCs typically favor RF and neural network models. In terms of performance, the RF model frequently yields the most favorable results. Furthermore, the RF model is capable of predicting molecular systems with higher PCE through extension, and its predictions can be validated through experimental methods, thereby demonstrating its accuracy in forecasting OSCs devices. For instance, Min et al.[58] applied an RF model to estimate and select appropriate D/A materials for synthesis and testing, achieving an actual PCE of 15.71%, aligning closely with the predicted values. Chen et al.[63] employed a trained RF model utilizing PM6 as a donor, which predicted a PCE exceeding 16% for five newly designed donor materials. The RF model's success stems from its capacity to effectively manage complex, nonlinear relationships within the data. It improves the accuracy and robustness of the model by building multiple DT and combining their predicted results. However, RF algorithms may exhibit a greater tendency to overfit compared to certain ML algorithms, such as Gaussian process regression (GPR), as RF can become overly attuned to the noise and intricacies present in the training data. Conversely, the GPR algorithm is not as flexible as RF when dealing with proxy functions with sudden jumps. Therefore, when employing a derivative ML model based on the RF algorithm to predict OSC devices, we can improve the model's generalization capability to new data while preserving accuracy by adjusting model complexity, utilizing regularization techniques, implementing data augmentation, integrating models, and applying CV. A compilation of models applied in ML studies for OSCs is presented in **Table 2**.

2.4. Self-Devised Machine Learning Predictive Models

The performance of OSCs is governed by a series of complex optoelectronic processes, including light absorption,[67] exciton

Table 2. A comparative overview of ML models of organic solar cells.

Acceptor	Data set	Sources	Model (optimal)	Performance	Year published	References
FAs	2.6 million	Literature	LR	$R^2 = 0.89$	2011	[51]
FAs	2.3 million	HCEP	Scharber	–	2014	[90]
FAs	5000	HCEP	CNN	AOC = 91.02%	2018	[91]
FAs	280	Literature	GB	$r = 0.79$	2018	[53]
FAs	1200	Literature	RF	$r = 0.62$	2018	[54]
FAs	300	Literature	GBRT	$r = 0.80$	2019	[98]
FAs	124	Literature	RF	$R^2 = 0.77$	2019	[100]
FAs	240	Literature	ANN	$r = 0.79$	2022	[99]
NFAs	565	Literature	BRT	$r = 0.71$	2020	[58]
NFAs	566	Literature	KNN	$r = 0.72$	2020	[129]
NFAs	154	Literature	RF	$r = 0.86$	2021	[59]
NFAs	265	Literature	RF	$r = 0.93$	2022	[60]
NFAs	717	Literature	XGBoost	$r = 0.79$	2022	[61]
NFAs	85	Literature	XGBoost	$R^2 = 0.86$	2023	[101]
NFAs	1242	Literature	RF	$r = 0.79$	2023	[62]
NFAs	397	Literature	RF	PCE > 16.00%	2023	[63]
NFAs	280	Literature	RF	PCE > 18.83%	2024	[64]
NFAs	310	Literature	RF	PCE = 16.24%	2024	[66]
NFAs	292	Literature	RF	$r = 0.81$	2024	[102]
NFAs	5000	Literature	RF	$R^2 = 0.92$	2024	[106]
NFAs	50 000	Literature	RF	$R^2 = 0.91$	2024	[68]

generation,[68] charge separation, and transport.[69–71] These processes are influenced by various factors such as the molecular structure of the organic materials,[72] the morphology of active layer,[73] and the interfacial properties between layers.[15,74,75] To enhance the efficiency and reliability of OSCs, researchers have been exploring the integration of ML techniques. Utilizing a self-devised ML framework, they have developed predictive models that demonstrate significant potential in accurately forecasting outcomes based on empirical data. The self-devised ML framework model is specifically tailored to address the unique challenges and complex working principles of OSCs. It includes a range of sophisticated ML algorithms aimed at analyzing and explaining the specific working principles of OSCs.

The exciton–drift–diffusion (XDD) modeling method presents a novel approach to examining the microscopic properties of thin film OSCs and predicting their optimal performance efficiency.[76] By reformulating the design problem as a surface optimization challenge and applying GA, the microstructure of the curves/surfaces is re-engineered. The XDD model shows a significant increase in current density in both 2D and 3D architectures compared to the traditional BHJ microstructure. Furthermore, using curve-based or surface-based methods to model the D/A interface significantly reduces the number of design variables. This reduction facilitates the use of complex gradient-free optimization techniques, essential for achieving peak performance in OSC design.

In 2023, Cole et al.[77] have developed a refined feature-selection method that combines a gradient boosting (GB) framework model with statistical feature analysis, mainly solving the problem of feature selection in high-dimensional data and improving the prediction accuracy and generalization ability of the model. Their comprehensive operational workflow is illustrated in **Figure 5** and is divided into four distinct stages: a) data acquisition and feature extraction; b) feature selection, analysis, and engineering; c) model optimization, and d) prediction of material properties. Their research positively impacted the field of feature selection, effectively reducing the computational burden during model training and prediction. It also offered effective strategies for addressing high-dimensional data challenges, which is essential for improving model interpretability.

The field of molecular architecture and synthetic methodologies is advancing rapidly, driven by the pursuit of superior organic semiconductors. However, empirically evaluating a wide range of organic compounds is costly. In response, Li et al.[78] have developed an innovative framework model that combines a deep learning model, specifically graph neural networks (GNN), with ensemble learning techniques like light gradient boosting machines (LightGBM). This integration marks a significant step in the efficient and accurate screening of OPV molecules. Notably, this specific framework model elucidates the intricate relationships between molecular structure, intrinsic properties, and device performance metrics and forecasts PCE based on the molecular physical and chemical traits. These predictions can be validated through experimental procedures. **Figure 6** in the study outlines their streamlined prediction workflow for FA OSCs, which consists of two main components:

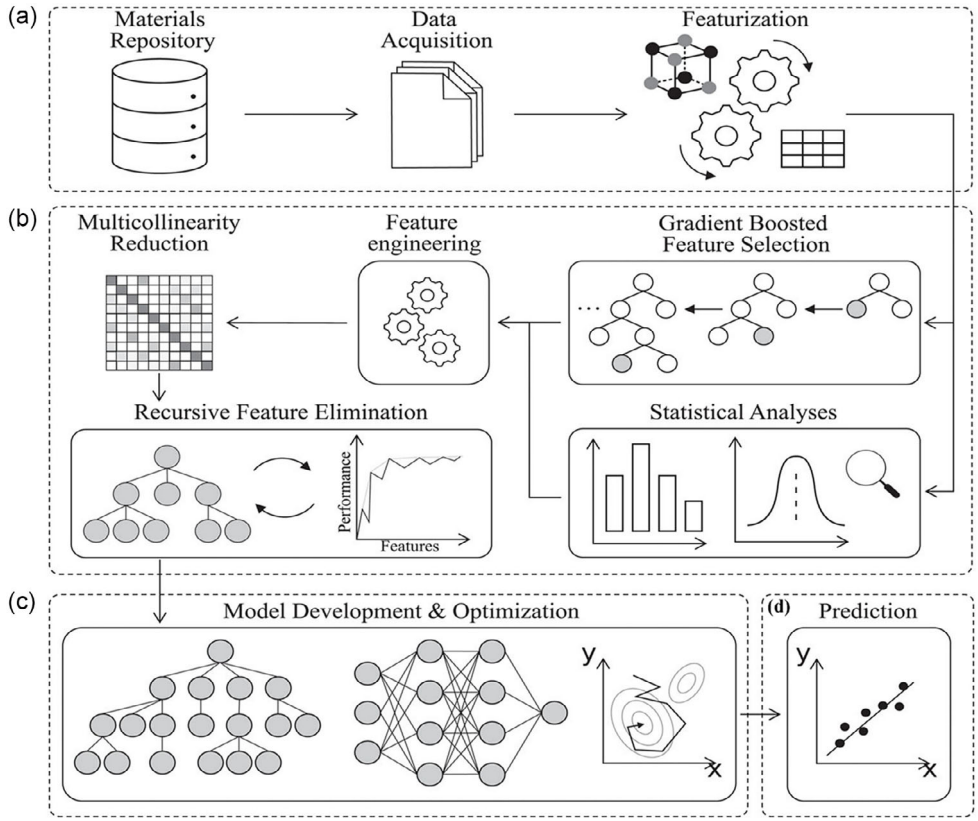

Figure 5. Overview of the operational workflow compartmentalized into four distinct stages: a) data acquisition and feature extraction, b) feature selection, analysis, and engineering, c) model optimization, and d) prediction of material properties. Reproduced with permission.[155] Copyright 2023, AIP Publishing.

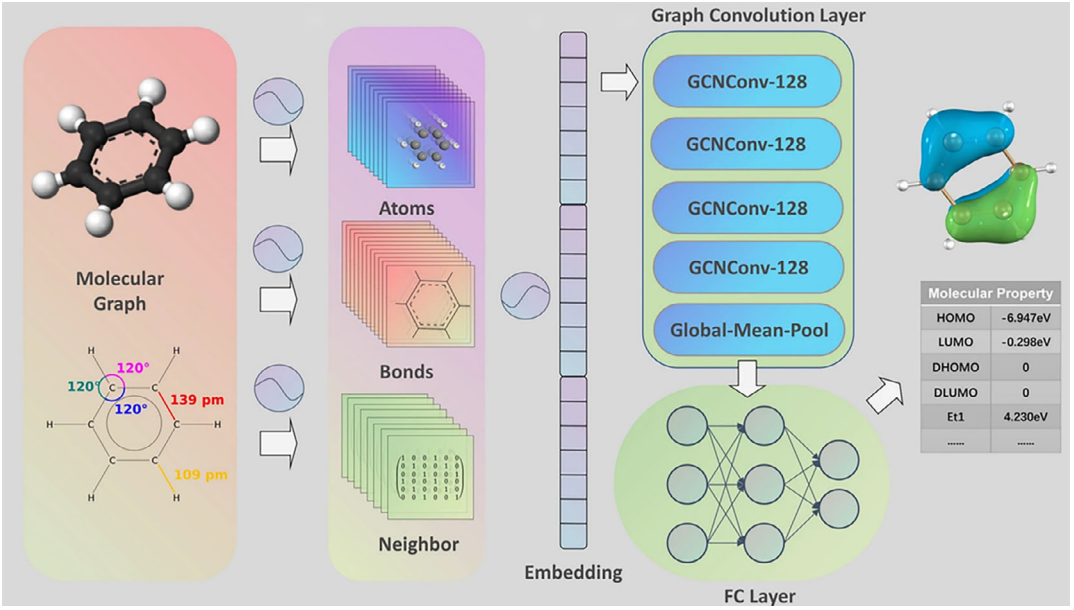

Figure 6. Streamlined prediction workflow for OSCs. Reproduced with permission.[78] Copyright 2023, Springer Nature.

1) using the GNN model to infer molecular characteristics from molecular topologies and 2) developing an ensemble learning algorithm to predict PCE based on the assessed physicochemical properties of the molecules. The result of this approach is a precise and nuanced evaluation of the chemical configurations of OPV molecules, which advances the frontiers of organic electronics research.

To clarify the complex relationship between the optical properties and chemical structure of organic materials, in 2024, Li et al.[79] crafted a predictive model to evaluate the optical bandgap of OPV materials. The methodology is illustrated in **Figure 7**. This model predicts the optical characteristics of a range of NFAs, including four newly synthesized materials for model validation. Based on the data transformation on the model, researchers found that strong noncovalent interactions exert a more pronounced effect on the band gap than weaker ones. Moreover, the addition of double bonds and asymmetry does not consistently result in a narrower band gap. These findings offer novel perspectives for the design and advancement of OSCs.

2.5. Automated Experimental Technology

Automated experimental technology is an essential tool for developing predictive models of OSCs. It enables researchers to conduct a greater number of experiments within a shorter period, generating extensive datasets that capture the nuances of OSC behavior under various conditions. This automation improves efficiency, reduces human error, and ensures more reliable and reproducible results.[80] Furthermore, the integration of AI has greatly advanced scientific research, especially in optimizing OSCs.[81] AI's ability to quickly and accurately analyze large datasets is now considered an essential resource.[82] Recently, Automated Machine Learning (AutoML) frameworks have been developed to bridge the gap in human expertise by building ML systems.[83]

Evaluating the industrial potential of OPV materials and devices requires multidimensional exploration within a vast parameter space. Manual experiments are limited in scalability and reproducibility. Automated platforms for manufacturing and characterizing functional devices can enhance experimental efficiency through precise process control. AMANDA, the Autonomous Materials and Device Application Platform, is a versatile platform for distributed materials research, featuring a self-developed software backbone and multiple Material Acceleration Platforms (MAP). The LineOne (L1) within AMANDA is tailored for the production and characterization of solution-treated thin-film devices such as OSCs. Du et al.[84] first demonstrated the evaluation of OPV materials and devices based on the high-throughput automation platform L1 in terms of efficiency and photostability. They automated the characterization of OPV materials and devices, employing GPR to predict photovoltaic parameters and aging losses with promising accuracy in **Figure 8**. The PM6: Y6 model material system was used to manufacture devices in air, achieving a maximum PCE of 14%.[85] Recently, they integrated automated experimental

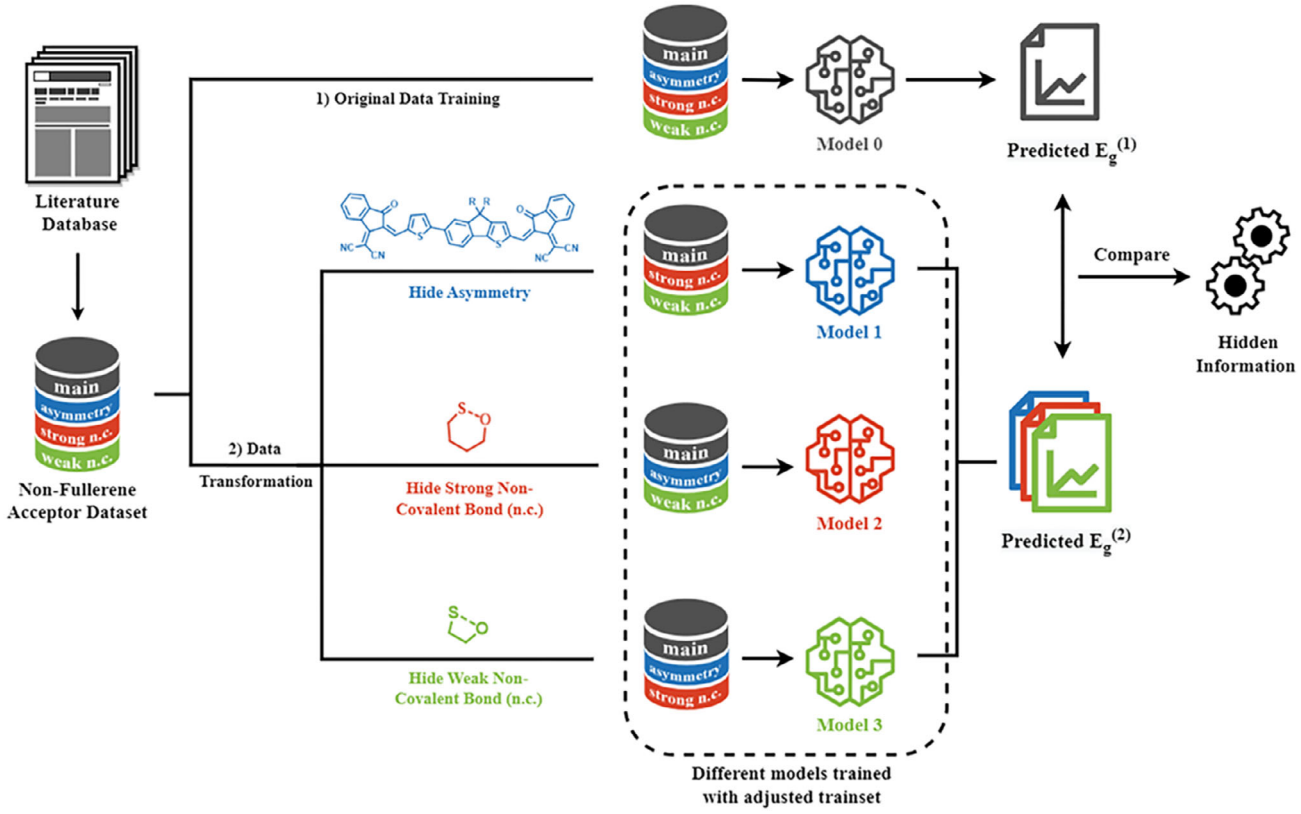

Figure 7. Flow diagram of optical bandgap prediction and explanation. Reproduced with permission.[79] Copyright 2024, John Wiley and Sons.

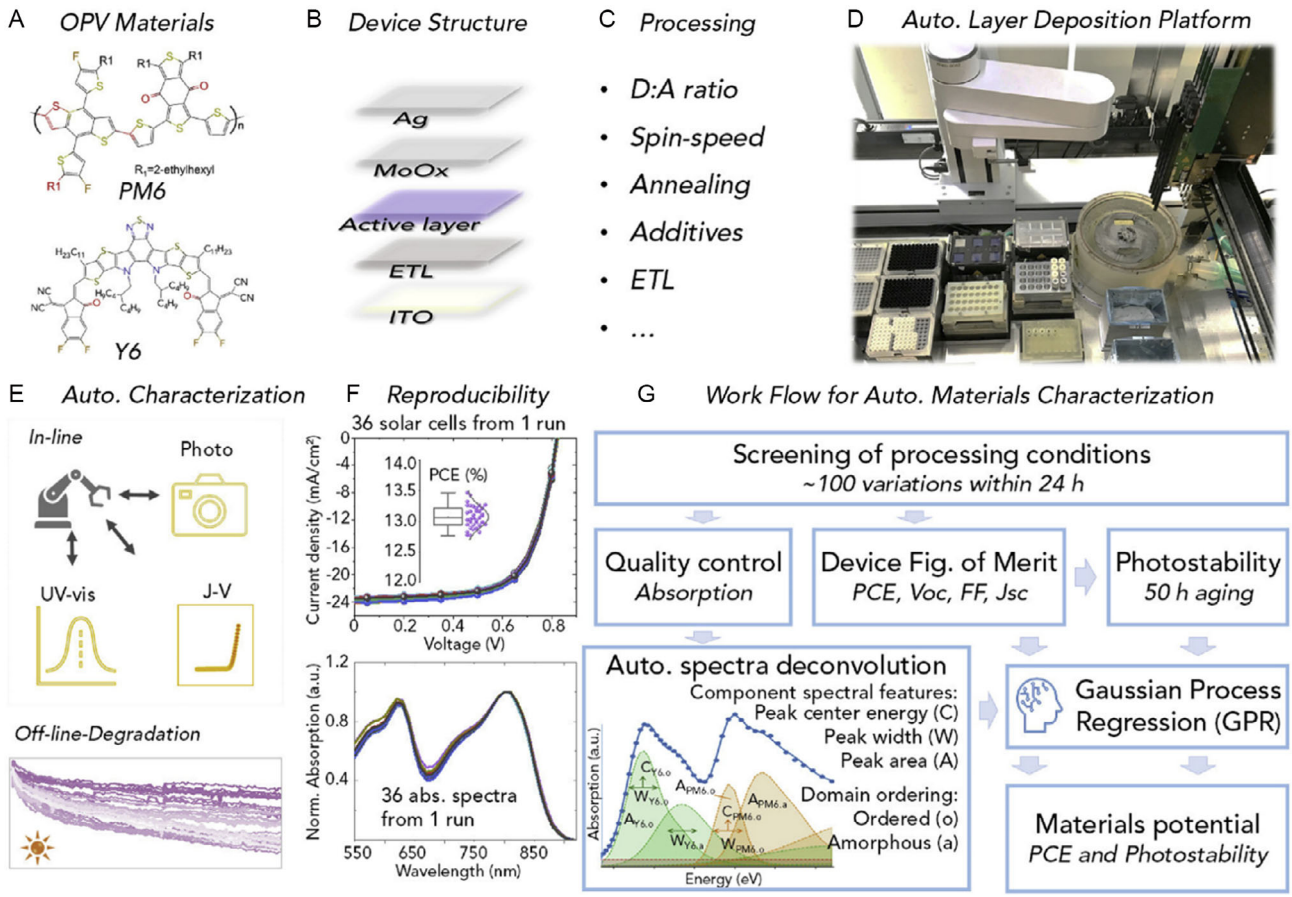

Figure 8. Based on the high-throughput automation platform and workflow: a) representative organic semiconductor materials, b) device structure, c) typical dimensions of processing variations for OSCs, d) photograph of the automated spin-coating layer deposition as part of the automated platform, e) schematic illustration of the automated in-line characterization and off-line photodegradation, f) processing reproducibility demonstration, and g) workflow for evaluating OPV materials in terms of efficiency and photostability with GPR-based data analysis. Reproduced with permission.[84] Copyright 2020, Elsevier.

technology with ML to screen for processing stability in various environmental conditions and indoor lighting, as well as to evaluate over 40 D/A combinations.[86] Utilizing AutoML technology, they aim to identify various materials and ascertain their structural characteristics that govern the resilience to light-induced degradation. By correlating electrical performance loss with spectral evolution, they have elucidated the stability of OPV devices with exceptional detail. They have pinpointed specific structural motifs responsible for the degradation of photovoltaic parameters and shown that air/light elasticity can be predicted through frontier orbital energy levels and active layer morphology, as deduced from UV-vis spectral analysis. This approach can help determine the most probable degradation mechanisms. Furthermore, they discovered that similar good predictions of air/photoelasticity can also be achieved by only considering the characteristics of chemical structure. The combination of high-throughput, superior precision in controlling experimental details and outstanding data quality will contribute to building digital material twins, which will then ultimately lead to the acceleration of materials science. Moreover, the ongoing development of automated experimental technology is expected to further enhance the efficiency and precision of research methodologies.

However, the main challenge in identifying ideal OSCs materials not only encompasses the selection of optimal algorithms and models but also includes database construction and descriptor selection. The following discussion will explore the detailed influence of databases and descriptors on the efficiency of OSC devices.

3. Databases

The ongoing progress in ML within the domain of OSC research has outpaced the capabilities of high-throughput computing databases, which now fail to satisfy researchers' requirements. These databases offer only the microscopic properties of materials without correlating actual device efficiencies. Therefore, researchers employ a variety of ML algorithms to uncover the inherent relationships between OSCs. They also build a comprehensive molecular structure and device performance database, which is grounded in experimental data. In this review, we assessed the research progress in OSC databases, distinguishing them into large-scale databases and small-scale experimental datasets, categorized by the volume of data points—above or

below a threshold of 10 000. Large-scale databases are rich in experimental data, providing a broader spectrum of information for researchers, whereas small-scale experimental datasets are ideal for in-depth study and validation of specific inquiries.

3.1. Large-Scale Databases

Large-scale databases often contain a wealth of experimental data, providing researchers with a more comprehensive understanding of training and validating ML models. Additionally, the abundance of data can speed up the comparison of different algorithms' performance, allowing researchers to quickly identify the most suitable algorithm for their specific needs. Essentially, large open-source databases are invaluable resources for researchers, aiding in the successful execution of their research projects.

Hutchison et al.[87] investigated over 90 000 π-conjugated copolymers, using a GA to predict their potential efficiency as organic heterojunction photovoltaic devices from the chemical structure space. They analyzed trends in copolymer sequences and key motifs in monomers and dimers to identify high-efficiency fullerene OSC materials. This was the largest database of conjugated polymers in the field of photovoltaics at that time.

Zhang et al.[88] created a database containing over 111 000 molecules using quantum chemical calculations. They evaluated various ML algorithms to identify the applicable range and established the applicability domain by considering the Euclidean distances to the training set. Employing an RF model, they efficiently predicted the HOMO and LUMO orbital energies determined by DFT B3LYP, achieving MAE of up to 0.15 and 0.16 eV, based solely on molecular descriptors related to connectivity. This method facilitates the accelerated screening of organic semiconductors for solar cell applications.

In relation to the OSCs, Harvard University has developed renowned open-source databases: The Harvard Clean Energy Project (HCEP), as shown in **Figure 9**a, was introduced by Aspuru-Guzik et al. They used DFT to extensively analyze 2.3 million virtual molecules for OPV.[89] This analysis yielded computational results for various molecular properties, including the energy levels of the HOMO and LUMO orbitals, energy gaps, and other relevant characteristics. Additionally, they employed the Scharber model to predict the PCE of these molecules.[90] The HCEP database, rich in data, serves as an excellent resource for developing a QSPR model. Utilizing deep ANN algorithms, accurate predictions of molecular structure performance can be made. Currently, neural network architectures are considered the most effective method for this purpose.[91]

As research on NFAs progresses, Aspuru-Guzik et al.[92] have conducted a statistical analysis of a database containing ≈51 000 data points, known as the NFAs Database (NFADB), as depicted in Figure 9b. The PCE of these materials was ascertained by calculating the calibrated energy levels of the HOMO and the LUMO. Predictions of the PCE for molecular devices and the screening of NFA materials were made using quantum chemical calculations and an enhanced Scharber model.

Peter et al.[93] introduced a novel dataset from the National Renewable Energy Laboratory (NREL) for OPV applications, encompassing ≈91 000 molecules represented by simplified molecular input line entry system (SMILES) strings. This database is distinguished by its inclusion of larger molecules, some with up to 200 atoms, and molecules with extrapolated

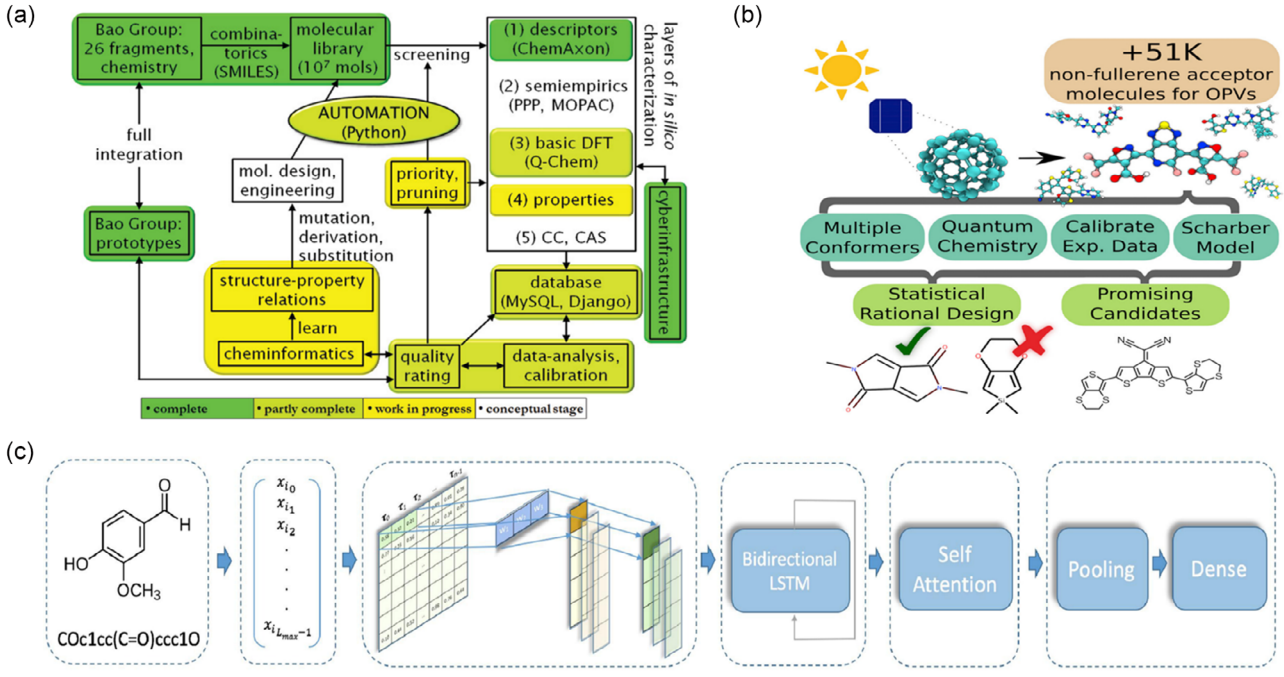

Figure 9. a) Structure and workflow of HCEP. Reproduced with permission.[89] Copyright 2011, American Chemical Society. b) Structure and workflow of NFADB. Reproduced with permission.[92] Copyright 2017, Elsevier. c) Detailed architecture diagram of the proposed attention-driven LSTM model. Reproduced with permission.[94] Copyright 2021, Elsevier.

properties of long polymer chains, which sets it apart from existing databases. The new database was subjected to photoelectric calculations for OPV applications using a message-passing neural network (MPNN).

Richards et al.[94] proposed an attention-driven long short-term memory (LSTM) network, leveraging both the HCEP and NREL OPV databases. They integrated natural language processing (NLP) derived algorithmic techniques into the OPV field. This approach, which strictly learns the SMILES representation of molecules, predicts photoelectric properties. For a more detailed understanding of the attention-driven LSTM model, refer to Figure 9c, which presents a comprehensive flowchart.

Recent research indicates that integrating multiple databases can create a high-throughput database conducive to experimental applications, thereby enhancing the predictive performance of OSCs. For example, Tsai et al.[95] proposed four ML models using XGBoost and ANN methods to enhance binary OSCs by incorporating a third component, thereby creating ternary OSCs to elevate PCE. They employed both experimental and DFT-derived HOMO and LUMO levels to conduct efficient high-throughput virtual screening (HTVS) of top candidates based on PCE. The HTVS leveraged two distinct latent databases: one with 429 413 uniquely recombined ternary OSCs systems derived from experimental data, and the other sourced from the HCEP database.

In summary, large-scale databases should be designed to accommodate the diverse and ever-evolving landscape of OSCs, ensuring that they remain relevant and up-to-date resources. Establishing comprehensive OSC databases is a crucial step towards advancing the field of OPV. However, creating and maintaining such databases is a significant challenge.

3.2. Small-Scale Experimental Datasets

To investigate the application of various ML algorithms in the realm of OSCs, researchers have gathered experimental data from the literature to create small-scale experimental datasets. These datasets aid in enhancing the understanding of OSCs materials' properties and offer accessible and valuable research insights. Small-scale experimental datasets can be analyzed on a case-by-case basis, addressing some of the challenges of creating large-scale databases.

Saeki et al.[96] explored the effect of the number of data points in the NFA-OPV database on the predictive accuracy of models. Their findings offer guidance for developing high-quality databases. As depicted in **Figure 10**a, the correlation coefficient r swiftly rose from 0.25 at 50 data points to 0.79 at 200 data points. Beyond 300 data points, the r value plateaued at 0.82–0.84, as illustrated by the diagonal line in Figure 10b, which represents the predicted versus experimental PCE. Consequently, it has been determined that ≈300 data points are adequate for the NFA-OPV database.

Aspuru-Guzik et al.[97] proposed the Harvard Organic Photovoltaic Datasets (HOPV15), which is based on the HCEP database. They compiled experimental data on OPV materials from the literature and enhanced it with quantum chemical calculations. The dataset includes both calculated and experimental values of the microscopic properties for 350 small molecule polymer donor materials. They introduced a new semiempirical

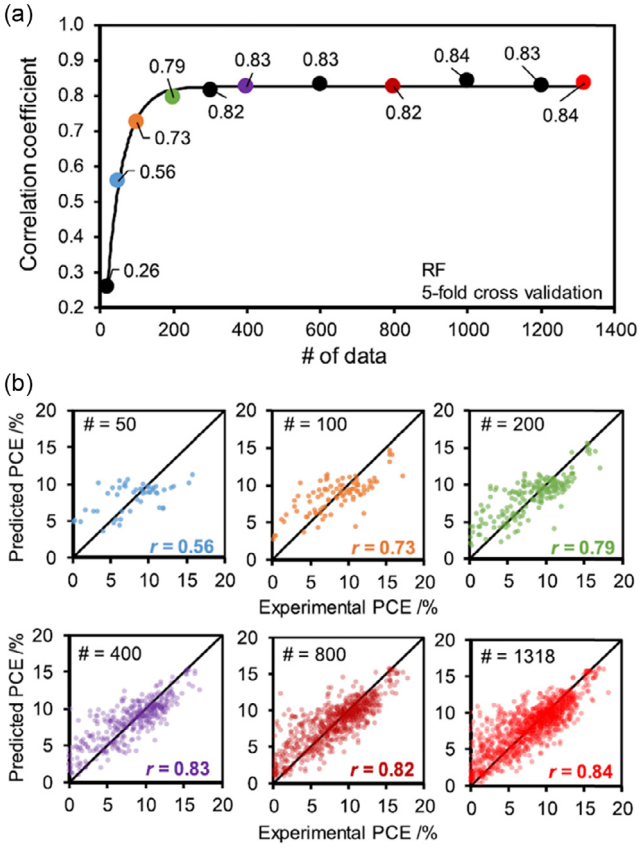

Figure 10. a) Effect of the number of data points of NFA-OPV on the correlation coefficient of RF model. b) The number of NFA-OPV data points affects the predicted PCE and experimental PCE (fivefold CV, r is the correlation coefficient). Reproduced with permission.[96] Copyright 2021, American Chemical Society.

method to generate a correction scheme that correlates electronic structure calculations with experimental data.

Ma et al.[53] constructed a dataset of 280 small molecule OPVs and employed various ML algorithms to predict the PCE model. They determined that the GB algorithm was optimal for predicting device PCE ($r = 0.79$). Subsequently, molecular descriptors were applied to an OPV dataset of 300 devices to predict V_{OC}, J_{SC}, FF, and PCE. The feature importance analysis function of GBRT was used to rank the importance of the molecular descriptors.[98] In 2022, Goharimanesh et al.[99] utilized 240 small molecule OSCs data from Ma et al.'s dataset.[53] They compared the performance of six ML models for predicting PCE. Expensive quantum chemical descriptors served as input, with ANN showing the best performance ($r = 0.79$) on the training set. The donor molecules, chosen following these guidelines, should be compatible with various device architectures or possess suitable NFAs to further enhance the performance of OPV.

Saeki et al.[54] collected experimental results from 1200 conjugated polymers, encompassing band gap, molecular weight, energy levels, and chemical structure fingerprints. They created datasets based on polymer molecular structure-device performance and utilized RF and ANN models for predictions. The RF model achieved a higher correlation coefficient ($r = 0.62$).

This work represents the first application of conjugated polymer material screening and design for BHJ OSCs.

Lee[100] collected 124 data points from the literature between 2012 and 2019 to create a dataset for ternary fullerene OSCs. He applied various ML algorithms to understand the complex relationship between the electronic properties of ternary OSC materials (conditional factors) and the device's photovoltaic parameters (target attributes). The RF model was identified as the optimal model for predicting PCE related to OSC devices, with R^2 of 0.66 and RMSE of 1.55. The CV was performed using the leave-one-out method, yielding an RMSE of 1.32, and the $k = 3$ resulted in R^2 of 0.77.

The scarcity of available datasets represents a significant challenge in establishing the structure-performance relationship of Y6-compatible donor molecules and in discovering new, efficient donor molecules. Zhou et al.[101] have developed a binary all-small-molecule OSCs (ASM-OSCs) database based on the Y6 of NFAs. They gathered 85 small molecular donors which matched with the Y6 from the relevant literature and employed four molecular structure descriptors to construct two ML models based on DT categorical boosting (CatBoost) and XGBoost. The goal of this study was to reveal the potential relationship between the donor molecular structure descriptors included in the models and the predicted PCE. These models, based on the XGBoost and CatBoost algorithms, demonstrated strong predictive performance, with R^2 values of 0.86 and 0.81, respectively.

Zhang et al.[63] have developed a comprehensive dataset of OSCs containing 397 D/A pairs, integrating photovoltaic parameters and descriptor sets. In 2024, Zhang et al.[64] collected data from 280 groups of D: A1: A2 type OSCs with NFAs to develop high-performance ternary OSCs. Then, they collected 292 D-NFA pairs with experimental OSC parameters from reported articles. Using the SHAP method, they analyzed the importance of the descriptors. Among the five ML models trained, the RF model showed the best performance ($r = 0.81$).[102]

Mariano et al.[103] have showcased the synergistic integration of high-throughput experimentation with statistical methods to efficiently screen and elucidate the photovoltaic potential of over 2000 low-bandgap OPV devices. In their evaluation of 24 D/A low-bandgap material combinations, featuring five distinct donors, the researchers discovered that blends based on PTB7-Th delivered high PCE. Moreover, electron affinity is a valuable descriptor for screening acceptor materials. Adjusting the chemical structure can alter the electron affinity, enabling the development of more efficient and application-specific electronic materials.[104] By chemically modifying the molecular quadrupole moment and fine-tuning the film morphology, as demonstrated by Kim et al.[105] the energy levels and recombination kinetics were optimized to enhance the V_{OC} of OPV while maintaining high charge generation efficiency. Electron affinity can be calculated extensively by combining DFT and molecular dynamics calculation to obtain electron affinity at different microstructures and compared with experimental results. Ahmad et al.[106] utilized a leading ML algorithm to predict the electron-affinity values of 5000 synthesized small molecules. Four ML models were employed and RF model demonstrated a high predictive capability, with R^2 values of 0.92 for the training set and 0.82 for the test set, respectively.

According to the above review of predicting OSC devices using small-scale datasets, it is evident that despite limited data, these datasets offer rich information when meticulously planned and analyzed. They can highlight subtle performance differences in OSCs under specific conditions. Moreover, compared to large databases, the generation and analysis of small-scale datasets require fewer resources. This efficiency is particularly advantageous for research groups with constrained budgets or limited computing power. Additionally, the flexibility of small-scale datasets can expedite the discovery process, fostering faster advancements in OSCs technology. Thus, creating targeted small-scale experimental datasets is not only practical but also strategic, offering a clear path to enhance our understanding and predictive capabilities of OSCs.

4. Descriptors

Descriptors, in addition to databases, play a crucial role in influencing model prediction accuracy. These descriptors encode the chemical and physical characteristics of molecules and serve as a vital communication tool for researchers working with databases. The task of selecting appropriate descriptors becomes notably difficult when the desired attribute is unclear.[107] In the field of organic photovoltaics, key prediction targets often involve optoelectronic material properties and OSC performance metrics, with input features naturally stemming from these molecular entities. These features typically encompass SMILES, molecular fingerprints, microscopic characteristics related to molecular structure and electronic properties, and weighted images of atoms and bonds.[108] The ML process depicted in **Figure 11** shows the prediction workflow for OSC performance, starting from chemical structure input and leading to the forecasted PCE. Therefore, a critical consideration when choosing descriptors is ensuring that the selected input features effectively and comprehensively capture the essential molecular properties right from the start.

This review has categorized the primary descriptor categories. However, the categorization is not entirely precise due to the use of various feature inputs in some studies.

4.1. Molecular Descriptors

Descriptors in ML are often vectors, matrices, integers, or other special character-based data structures, used for predicting molecular properties.[109] They are divided into experimental and theoretical categories. However, no single representation method is universally applicable for all properties. Molecular descriptors cover a range of features, from simple, such as the count of specific atoms, to complex, such as charge distributions. Zero-dimensional (0D) descriptors, known as count descriptors, provide information about the number of atoms and molecular weight, not structure or connectivity. 1D descriptors list structural fragments like functional groups, with fingerprints being a common example. 2D descriptors offer information based on the graphical representation and bonding of atoms. 3D descriptors, the geometric descriptors, describe the size, surface, and spatial coordinates of atoms in molecules. Four-dimensional (4D) descriptors, also grid-based, add a fourth

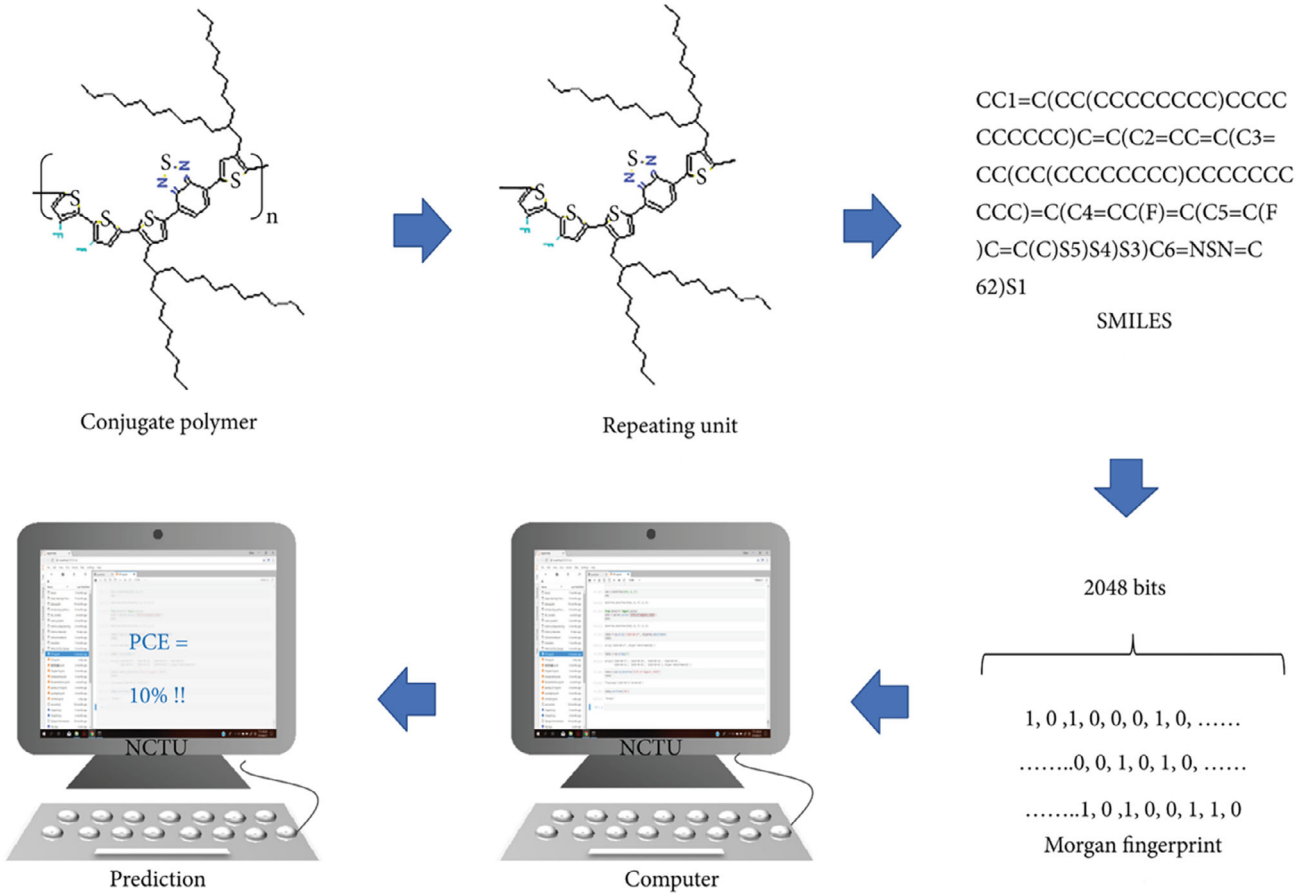

Figure 11. Schematic workflow of the machine learning for predicting the performance of conjugated polymers in OPV devices. Reproduced with permission.[156] Copyright 2019, John Wiley and Sons.

dimension to molecular geometry, typically describing interactions with an acceptor's active site or various conformational states.[110]

The use of SMILES as a feature input has gained significant attention in recent years, particularly following the development of 1D text encoding algorithms. More complex molecular structures are often represented through weighted graphs.[111–113] SMILES can be easily imported into various molecular editing software programs, which can then generate 2D graphics or 3D molecular models. Paul et al.[114] introduced a transfer learning method from the HCEP database to two smaller datasets within the HOPV15 database. The DNN architecture, illustrated in **Figure 12**, consists of two branches: one for SMILES and one for International Chemical Identifier (InChI) input types. Here, SMILES represents the primary structural type, while InChI indicates protonation states. The objective of SINet is to learn distinct molecular representations captured by different encodings. Aspuru-Guzik et al.[97] also applied transfer learning to predict HOMO values in the HOPV15 database based on predictions from the HCEP database. The computational complexity associated with using 3D geometric shapes as descriptors in neural network models hinders the widespread application of ML methods in high-throughput screening. Peter et al.[93] developed a DFT photoelectric computational model tailored for OSC applications, leveraging MPNN that begin with the SMILES descriptor to forecast molecular properties based on the 2D structure of molecules. This approach can yield nearly optimal prediction results for larger molecules with practical applications, eliminating the need for optimized 3D geometric shapes as input.

Jørgensen et al.[115] utilized an unsupervised learning VAE to predict molecular properties and generate novel molecules with specific characteristics, utilizing predefined SMILES. Alperstein et al.[116] improved the VAE method by integrating a more efficient RNN to encode multiple SMILES of a single molecule in a stacked fashion. This strategy combined SMILES representations from different inputs and utilized an attention-based aggregation approach to create the final latent representation. Through translating molecules into a set of nonoverlapping SMILES, All SMILES VAE technique established an almost objective mapping that connects molecules to existing latent representations within a high-probability density subspace.

Saqib et al.[117] have predicted the reorganization energy of OPV using ML models. They employed statistical analysis to select the best molecular descriptors. Ten ML models were developed and evaluated to identify the one with the highest predictive

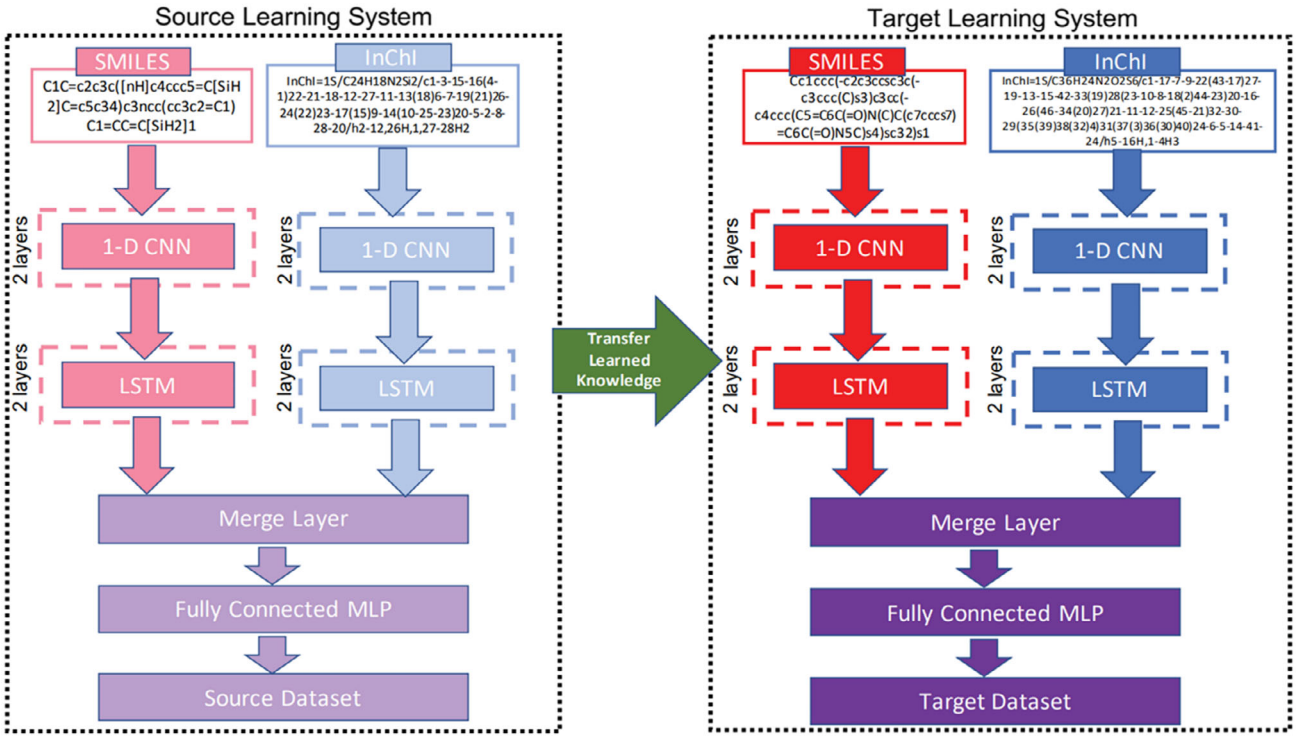

Figure 12. The proposed SINet architecture for learning from the two text-based molecular representations: SMILES and InChI. Reproduced with permission.[157] Copyright 2019, IEEE.

accuracy. The molecular descriptors were calculated and subjected to a thorough statistical and visual analysis. New organic semiconductors were designed, and their reorganization energies were predicted. The most promising candidates for photodetector applications were selected, and their energy levels were calculated using quantum chemical methods. The overall framework is depicted in **Figure 13**.

Using SMILES-based molecular representations has shown significant benefits in ML tasks. These representations are able to be transformed into RDKit molecular objects, which allow for the extraction of structural data and comparison of similarities. In a study by Janjua et al.[65] they explored chemical similarity analysis with RDKit descriptors, aiding in visualizing chemical space, evaluating structural diversity, and assessing monomer similarities. This method integrates ML, cluster analysis, and synthetic accessibility, offering a structured framework for designing polymers in OSCs. Karak et al.[62] combined frontier molecular orbital and RDKit descriptors to enhance the predictability of a RF model for PCE.

In contrast, the GB model exhibited notably improved performance in predicting V_{OC} and J_{SC}, achieving high r of 0.842 and 0.862, respectively. This success highlights the effectiveness of merging molecular and microdescriptors for strong correlation. By incorporating a variety of descriptor types, prediction accuracy can be enhanced. With the guidance of these innovative descriptors, the ML framework is positioned to simplify the development of new molecules and streamline the screening and prediction of suitable D/A pairs, ultimately expediting the advancement of efficient OSCs.

4.2. Molecular Fingerprint Descriptors

Digital representations of chemical structures, known as molecular fingerprints, are created by specific calculation rules, converting them into fixed-length vectors of bits. The length and information content of these fingerprints vary based on the calculation approach. By enabling efficient database searches, they allow for the comparison of molecular similarities. Their benefits encompass a uniform bit length and ease of acquisition, making them ideal inputs for ML and high-throughput screening applications.[118]

In 1980, the idea of utilizing molecular fingerprints to characterize molecular structures was introduced for the first time.[119] Aspuru-Guzik et al.[120] harnessed the extensive HCEP database and employed molecular fingerprints within a multilayer perceptron (MLP) model to forecast the HOMO and LUMO energy levels of molecules and the PCE of devices, achieving MAE of 0.028 and 0.032 eV, respectively. Min et al.[58] divided D/A molecules into fragments, generating molecular fingerprints from ASCII code strings, and applied ML models for screening, which demonstrated remarkable predictive performance. **Figure 14**a shows the conversion process of D/A pairs' chemical structures into digital data. Sun et al. created an automated design system named La FREMD, which integrates fragment molecular fingerprints with ML algorithms by segmenting molecules. This system, depicted in Figure 14b, consists of four distinct components.[121] The initial component uses a proprietary database alongside the La FREMD fingerprint map to evaluate the relevance of 6180 unique organic molecular

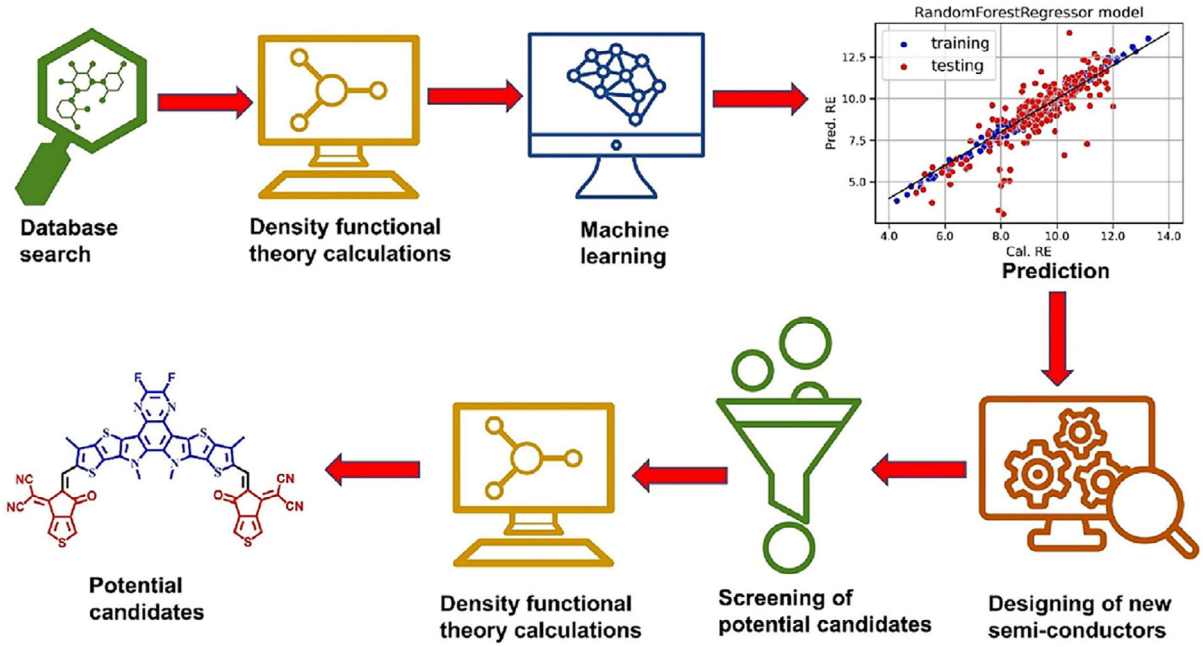

Figure 13. Complete framework (different steps involved in ML analysis) of the current study. Reproduced with permission.[117] Copyright 2024, Elsevier.

fragments. This feature engineering step identifies critical basic units that significantly impact performance. The second component generates a virtual library of new materials by reorganizing the most important building blocks on two promising backbones. The third component utilizes various ML algorithms such as ANN, RF, SVM, and GBRT to train and enhance ML regression models for the OSC donor material database. The most effective ML model is then employed for new material screening. The final component further predicts the potential PCE of devices based on selected donor candidates and the acceptor Y6. This research enhances the accuracy of ML model predictions by extracting interpretable structure–activity relationships for OSCs, offering a novel approach to proactively and spontaneously design materials rather than merely passively screening them.

Saeki et al.[54] utilized the Molecular Access System (MACCS) and Extended Connectivity Fingerprint (ECFP6) to train ML models, reaching an accuracy of 48% in forecasting PCE values using an RF model. They then fabricated new polymer-fullerene OPV devices with a PCE of 0.53%, significantly lower than the RF model's projected range of 5.0–5.8%. The ANN, known for their proficiency in recognizing images and language, demand abundant data for constructing accurate models. Due to restricted data availability and a large number of independent input factors, the RF model's correlation coefficient surpasses that of ANN by a significant margin. Building upon the work of Saeki and team, Chen[122] leveraged SMILES strings and average PCE to estimate device PCE values by generating chemical structure fingerprints using the Morgan algorithm. This strategy offers an advantage in predicting the PCE of potential conjugated polymers without requiring prior knowledge of material properties. The research also evaluated the efficacy of SVM and RF models, achieving a strong correlation coefficient of $r = 0.653 \pm 0.015$ through a combined approach known as an ensemble method. This amalgamated learning technique enhances the accuracy of PCE predictions derived from chemical structure fingerprints, opening up a novel pathway for the rapid screening of conjugated polymers in OSCs.

Recently, Hu et al.[68] developed a comprehensive library of NFAs. They trained various regression models, including bagging regression (BR), RF, GB, and LR, to predict exciton binding energies. The RF model emerged as the most accurate, with R^2 value of 0.91. The Breaking Retrosynthetically Interesting Chemical Substructures (BRICS) methodology[123] was employed for the design of novel NFAs. The predicted values guided the selection of these NFAs. Subsequently, clustering and chemical similarity analyses were conducted on the selected NFAs using chemical fingerprints. Additionally, the synthetic accessibility of the new NFAs was also evaluated.

4.3. Microscopic Descriptors

Although molecular descriptors and molecular fingerprint descriptors are quick and simple to compute, they do not provide accurate predictions of OSC performance. On the other hand, microscopic descriptors offer a more comprehensive range of molecular characteristics that are closely linked to the process of converting light into electricity. These descriptors in the field of OSCs consist of properties such as the optical gap, thin film morphology, carrier mobility, binding energy between holes and electrons, and the energy levels of molecular HOMO/LUMO. They have proven to be reliable in forecasting the performance of OSC devices. Nonetheless, these descriptors come with limitations. Quantum chemical computations, which are essential for determining the 3D structure of molecules and conducting

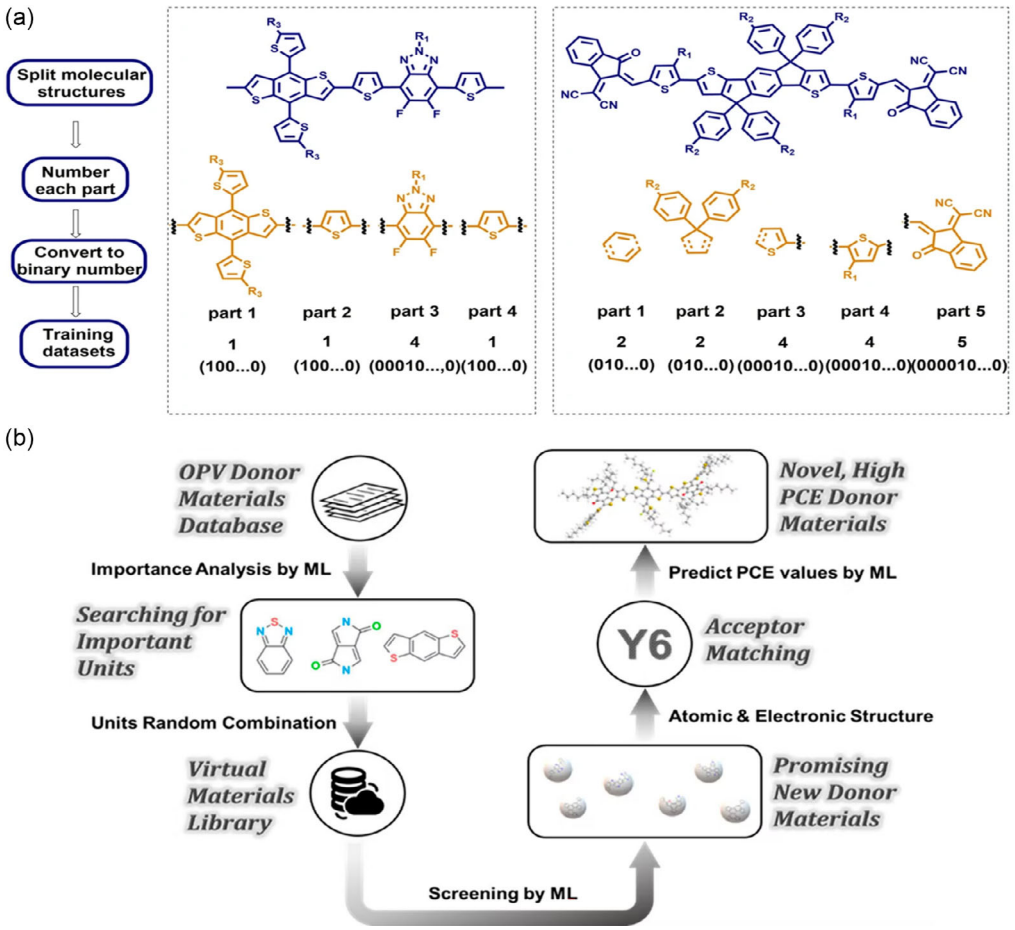

Figure 14. a) Diagram for translating chemical structures of D/A pairs into digitized data. Reproduced with permission.[58] Copyright 2020, Springer Nature. b) Scheme of an AI design framework for developing high-performance OPV donor materials. Reproduced with permission.[121] Copyright 2021, American Chemical Society.

analyses in the most stable configuration, are time-intensive and not suitable for large-scale material screening.

Aspuru-Guzik et al.[124] introduced a new technique for adjusting quantum chemical properties by utilizing the HOPV15 database. GPR was utilized to enhance electronic characteristics, encompassing frontier orbital energy levels, optical gaps, and key device parameters like V_{OC}, J_{SC}, and PCE. Subsequently, this approach was expanded to the NFADB database to fine-tune the calculated HOMO and LUMO energy levels, streamlining the evaluation of NFAs.[125] Lee's innovation involved constructing an RF model by leveraging existing literature data to systematically evaluate the HOMO and LUMO energy levels and their respective disparities in OSCs molecules. This model was then utilized to forecast the efficiency of binary,[126] ternary,[100,127] and tandem[128] OSC devices. Moreover, Lee conducted an investigation into the relationship between the molecular energy levels of D/A pairs and their importance features to gain deeper insights into the underlying mechanisms. Moreover, Ma et al.[53] developed a predictive model for PCE utilizing 13 microfeatures as descriptors, attaining an RMSE of 1.07 and r of 0.79 through the application of the GB algorithm. Nevertheless, the computational burden associated with these descriptors, such as polarizability and excitation state, could be substantial, potentially impeding large-scale high-throughput virtual screening processes.

Trois et al.[129] compiled a dataset consisting of 566 D/A pairs, focused on investigating two categories of descriptors for ML purposes: structural (which includes topological properties) and physical (encompassing molecular energy levels, molecular size, light absorption, and mixing characteristics) to predict PCE values. It highlights the relative importance of these descriptors, demonstrating the RMSE optimization results for the KNN algorithm with hyperparameters $\gamma_1 = 0.65$, $\gamma_2 = 4.08$, and $\gamma_3 = 0.97$. These hyperparameters indicate that the donor's fingerprint is the predominant factor influencing prediction accuracy, while the acceptor's fingerprint and five physical descriptors also have a significant but comparatively smaller impact. For a large database of hypothetical molecules, the initial screening should prioritize structural fingerprints before conducting a mid-level accuracy screen to identify the most promising candidates. Hence, when introducing a new compound into an ML model, it is essential to preserve a robust selection of physical descriptors to maintain high predictive accuracy.

Zhao et al.[130] utilized ML predictive models to investigate the intricate connections between structures of D/A, electronic

characteristics, and nonradiative voltage loss ($\Delta V_{OC}^{non-rad}$) in NFAs. The current literature lacks studies on the impact of frontier molecular orbitals, energy level differences, and optical band gaps of D/A materials on $\Delta V_{OC}^{non-rad}$ in NFAs. This research established an extensive model that considers building dataset, feature engineering, ML, and model evaluation, as illustrated in **Figure 15**.

To analyze the relationship between algorithm efficiency and material attributes, four ML algorithms—KNN, support vector regression (SVR), RF, and XGBoost—were employed to construct the forecasting model. By adjusting the parameters and hyperparameters of the algorithms, the model's generalization error is minimized. The model's predictive capability is assessed using metrics such as R^2, RMSE, and MAE, with the top-performing model chosen for further forecasting.

Zhang et al.[102] collected experimental articles on NFA-based OSCs, extracting and screening 292 OSC sample data with molecular structures, energy levels, and morphological characterizations, along with photovoltaic performance parameters. The descriptor set for training the ML model included the D/A mass ratio in the OSCs active layer, root mean square roughness, active layer thickness, maximum absorption wavelength of acceptors, energy levels, and molecular structures. They performed a feature importance analysis using the SHAP method to assess the impact of descriptors on the target property and PCE as shown in **Figure 16**. The model evaluation revealed that the RF model exhibited superior predictive ability and stability, with r of 0.78 for the training set and 0.81 for the test set.

4.4. Image Descriptors

The depiction of molecules in a molecular image naturally conveys their structure. Adhering to the empirical rules of bond formation, molecules can be visualized as undirected graphs where atoms act as "nodes" and bonds serve as "edges." This image-based visualization comes with several advantages: it presents the 3D architecture of molecules more straightforwardly compared to string-based formats, which rely on specific rules or algorithms to convert molecular structures into sequences of characters.[131] This graphical representation offers clearer insights, making it easier for researchers to comprehend the structure and properties of molecules.[132] By examining 3D structural images, one can swiftly understand key aspects of a molecule, including its shape, dimensions, bond angles, and other vital features.

ML has made significant strides in the field of image identification through the utilization of algorithms like CNN and attention mechanisms. These tools aid in extracting detailed feature data, thereby enhancing the precision and dependability of forecasts. In research conducted by Zhao et al.[57] CNN were utilized to create and assess models for NFA generation. These models included the use of dilated convolutional layers for feature extraction, along with an interpretable attention mechanism. This investigation showcased the superior performance of image representations when compared to string representations.

Sun et al.[91] developed an ML model called residual neural network (ResNet), which was based on a deep neural network utilizing the HCEP database, as shown in **Figure 17**. The model used unprocessed images of molecular structures as inputs, extracting important features from the structures and automatically linking them to PCE. The model achieved an impressive prediction accuracy of 91.02% for donor materials. This study directly predicts the photovoltaic performance based on the chemical structures of various donor materials for OSCs, requiring minimal computational resources and speeding up the screening of new OSCs. Subsequently, Sun et al.[133] explored different representations of molecular structures, such as images, ASCII strings, and molecular fingerprints. These representations

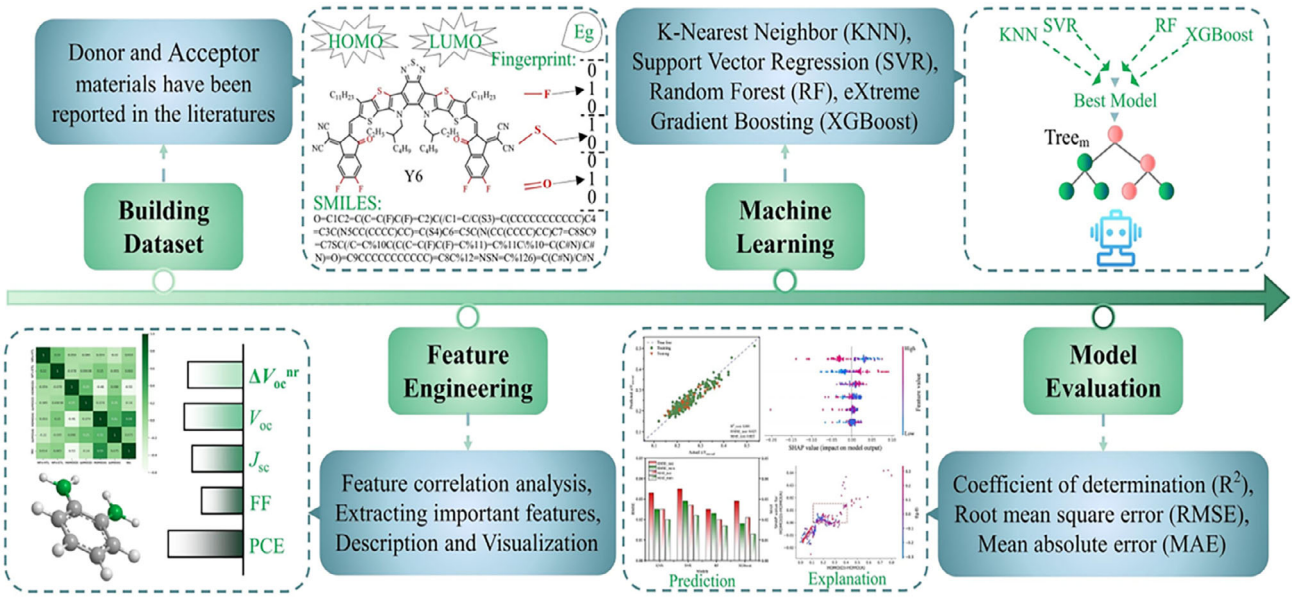

Figure 15. The flow chart illustrates the process of constructing a dataset, feature engineering, model selection, model evaluation, interpretation, and application. Reproduced with permission.[130] Copyright 2023, Elsevier.

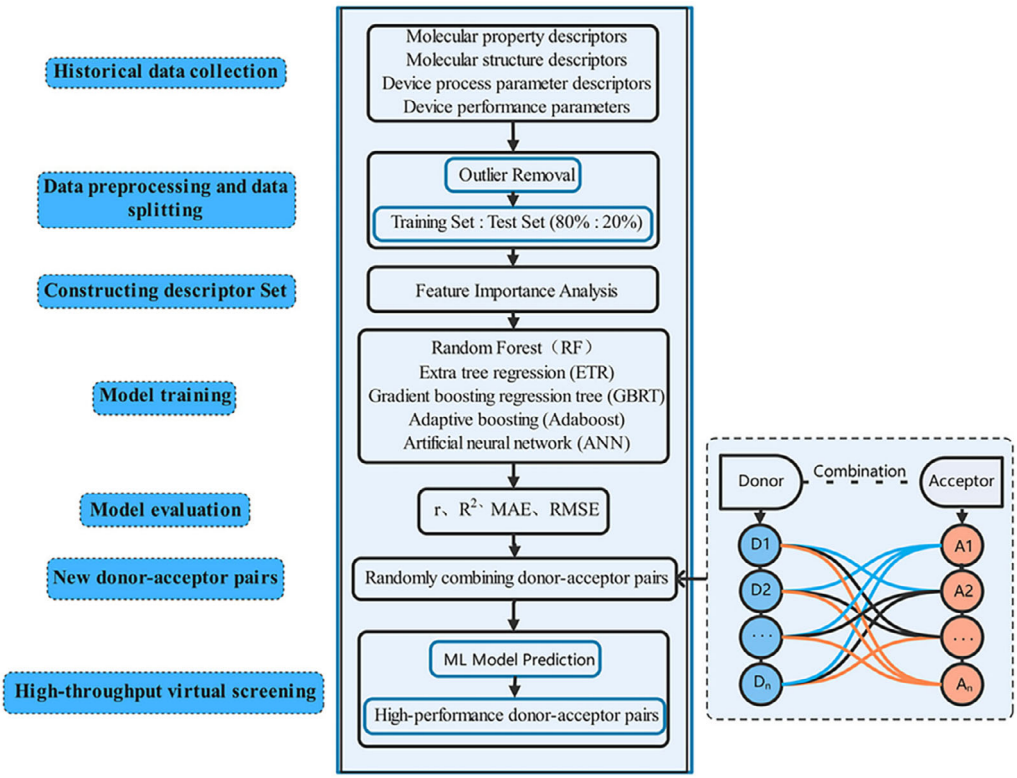

Figure 16. Scheme of model design framework for developing high-performance OSCs D/A pairs. Reproduced with permission.[102] Copyright 2024, John Wiley and Sons.

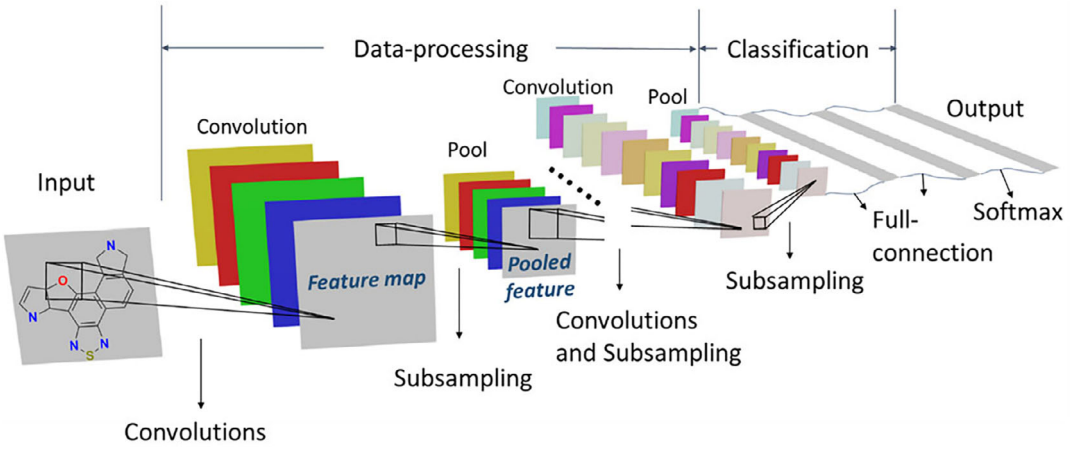

Figure 17. Predicting the PCE of OPV devices using CNN and molecular picture input. Reproduced with permission.[91] Copyright 2018, John Wiley and Sons.

were fed into various ML algorithms to establish relationships between structure and performance, with molecular fingerprints consisting of around 1000 bits showing the highest prediction accuracy.

In 2023, Li et al.[78] introduced a novel method for predicting molecular properties using GNN models, termed the "Property Model." This model, in conjunction with the LightGBM algorithm, known as the "Efficiency Model" was employed to predict the overall PCE of OSCs. The Property Model captures the relationship between molecular structure and properties, offering a valuable tool for chemical system analysis. As a deep learning architecture, GNN has demonstrated exceptional adaptability in processing molecular structures, facilitating its extensive application in chemical research. The frameworks of the Property and Efficiency Models are depicted in **Figure 18**.

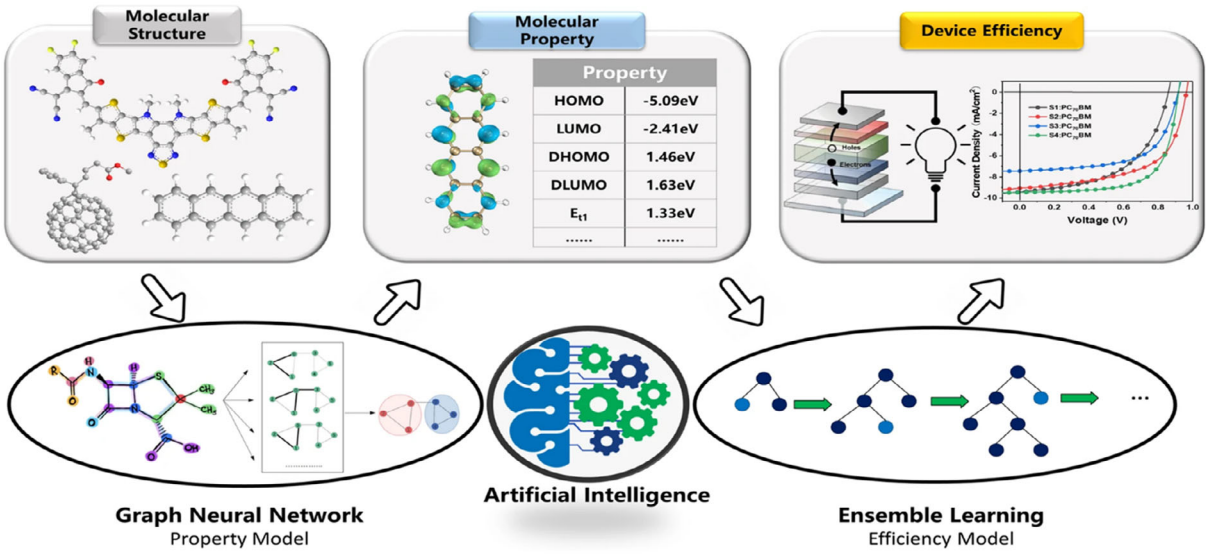

Figure 18. The framework of the Property Model and Efficiency Model. Efficiency Model predicts the final PCE of the OSCs. Reproduced with permission.[78] Copyright 2023, Springer Nature.

Martin et al.[134] curated a new dataset of OPV device performance, molecular structures, and a complete set of device fabrication parameters, totaling 558 data points. They linked each D/A materials with a molecular structure, represented as a SMILES string, as shown in **Figure 19**. In their study, extended-connectivity fingerprint efficiently captured local structural information, and the polymer-unit fingerprint was introduced to offer a more interpretable fingerprint-based representation for conjugated molecules. The histogram gradient boosting (HGB) model, a gradient-boosted DT algorithm, was identified as the most effective for predicting PCE from molecular structure alone and in combination with device fabrication features. Even with modest R^2 values of ≈0.6, this represents the cutting-edge predictive performance for PCE.

Images are not the sole representation of data, a fact that can be advantageous for data augmentation but may also be disadvantageous when used as model inputs due to the potential introduction of noise. Consequently, there is scant research on employing images as inputs for ML models in OSC devices.

5. Challenges and Prospects

The molecular structures of OSCs materials are intricate and varied, positioning ML as an inherently appropriate technique for the OSCs field. ML approaches have demonstrated significant potential in the design of OSCs, yet their role in fostering innovation and breakthrough discoveries remains constrained.

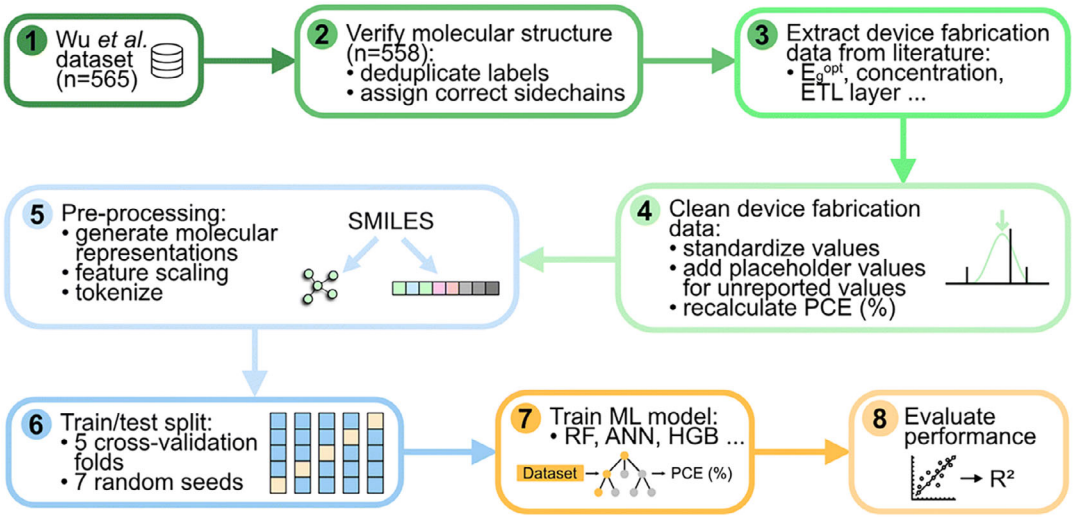

Figure 19. Flowchart describing the steps involved in curating the first dataset of OPV device fabrication data, and testing and validation of representations and models. Reproduced with permission.[134] Copyright 2024, Royal Society of Chemistry.

Typically, ML is only optimized within the given search space, and human intuition and creativity are still crucial for achieving groundbreaking innovations. ML can augment the human creative process by visualizing feature spaces, which assists researchers in uncovering new patterns and trends within complex data, thus sparking fresh ideas and solutions. The application of ML approaches to OSCs encounters various challenges, including developing accurate predictive models, sustaining extensive and varied datasets, and meticulously choosing descriptors to enhance the accuracy of efficiency forecasts. The subsequent section offers valuable insights into these challenges, which should aid researchers in the OSCs field.

5.1. Predictive Model Construction

Structure–activity relationship models have yielded some findings in the OSC domain, yet they confront ongoing challenges.[135] Quantum chemistry-based function predictive models provide precise results but require significant expertise and computational power, which limits their use in high-throughput screening processes. Moreover, the intricate microscopic environment of compounds often leads to descriptors that omit crucial chemical details, thus yielding suboptimal predictions. Currently, ML models are widely used for material screening and design in the field of OSCs.[136–138] However, these models are often considered "black boxes," making it difficult to interpret their results. The Scharber model and its extensions can clarify the impact of various parameters on OSCs but tend to neglect the molecular properties at a microscopic level. For example, Saeki et al.'s attempt to predict new polymer OSCs using the Scharber model was not successful, due to its foundational limitations and expansion obstacles,[54] which hinder the inclusion of descriptors like structure, topology, and thermodynamics.[129] In contrast, Troisi et al.[139] have shown that ML models trained with electronic and structural parameters can perform better. Additionally, the Scharber model's reliance on the "older" FA PCBM is less suitable for high-performance OSC material research.[140] On a positive note, Hutchison et al.[50] introduced the OPEP model for predicting NFAs, which significantly speeds up computations and efficiently screens for new, efficient NFAs.

We have observed an interesting trend: despite differences in data volume and feature inputs, the majority of studies have utilized RF models, yielding impressive results. ML models, such as GB, XGBoost, and deep learning, which are built upon RF models, could be promising directions for future research in OSCs. These models enhance RF capabilities by employing more advanced algorithms and larger datasets, thus improving prediction accuracy and optimization results. As the ML approach matures and OSC research advances, derivative models of RF are likely to play a crucial role in driving future studies.

5.2. Database Quality and Quantity

Open-source databases such as HCEP, HOPV15, and NFADB serve as extensive resources for OSC research, akin to Wikipedia. Realizing this vision requires a robust OSCs data repository, encompassing both historical and current device data, necessitating the engagement of the entire OSCs research community. The labor-intensive nature of DFT calculations presents a challenge for including extensive data in ML model training. Therefore, we recommend conducting a case-by-case analysis of specific issues, supported by the creation of specialized, small-scale experimental datasets. This targeted approach will guide the construction of databases using ML techniques. From an algorithmic perspective, leveraging meta-learning and Bayesian algorithms is a promising strategy.[141] Meta-learning facilitates knowledge extraction from limited data, while Bayesian algorithms balance data scarcity with model predictive accuracy. Constructing small-scale, targeted databases to address specific issues is a pragmatic approach in database development. This method is prevalent among researchers, as illustrated by the assembly of experimental data from literature to create specialized datasets.[142] Saeki et al. emphasized the impact of data quantity on ML model accuracy, indicating that ≈300 entries in the NFAs-OPV database suffice for reliable predictions.[96] The relationship between model accuracy and flexibility is complex; data quantity is not the sole determinant of model accuracy. These insights are instrumental for establishing high-quality databases focused on testing specific hypotheses.

5.3. Selection of Descriptors

While molecular descriptors and molecular fingerprint descriptors offer quick and simple calculations, they may not be sufficient for modeling OSCs. On the other hand, microdescriptors, which are closely linked to the process of converting light into electricity, can provide better predictive results. However, their use requires precise quantum chemical calculations that can be costly and impractical for large-scale screening due to computational constraints. As a result, researchers need to find a balance between computational speed and predictive precision.[141] This necessity has led to a need for the development of new descriptors tailored specifically for OSCs. The crucial factors influencing the efficiency of OSC performance include the materials in the active layer, solubility, and the inclusion of solvent additives.[143–145] Yao et al.[146] introduced a basic set of quantum chemical descriptors and developed a versatile predictive model that can accurately predict material solubility in various solvents. This work contributes significantly to the synthesis of OSC materials and the optimization of chemical processes, providing valuable insights into predicting high-performance OSC devices. They further utilized the RF algorithm with a concise set of descriptors, each capturing 7 bits of information.[147] These descriptors include the area (molecular surface area), σ^2+ (positive electrostatic potential energy variance when analyzing the distribution of electrostatic potential energy on a vdW surface), and σ^2- (negative electrostatic potential energy variance when analyzing the distribution of ESP on the vdW surface) of both the solvent and solute, along with temperature.

Currently, multiple studies have explored different experimental methods that combine various types of descriptors, such as molecular[62] and microscopic or a combination of molecular fingerprints with microscopic descriptors,[148,149] or incorporating descriptors specific to the device. Such descriptors are crucial for ensuring that models do not develop biases towards specific

aspects of the data, allowing for a more accurate understanding of how compounds behave in various environments. Employing a variety of descriptors can overcome the limitations inherent in relying on a single type. Certain descriptors may excel under specific conditions or for particular prediction tasks. By combining them, a model can capitalize on the strengths of each descriptor while mitigating the weaknesses of others. Ma et al. introduced novel device descriptors, namely, the D/A weight ratio and the root mean square of surface roughness, into the ML model, discovering that refining device parameters can significantly enhance OSCs efficiency.[150] Hence, the integration of multiple descriptor types can augment the predictive precision of ML models, offering direction for the future design of applicable descriptors.

5.4. Combining Automatic Experimental Technology with ML

Experimental validation is essential to confirm the predictions made by ML models. Extensive research has shown that ML models can predict the performance of OSCs with high accuracy within known datasets. However, when predicting novel materials, discrepancies often arise.[151] This is often due to the omission of various experimental factors such as the D/A ratio, stirring speed, solvent type, and processing temperature during the model training phase. There remains much to be understood about how environmental stresses, like temperature, humidity, and oxygen, affect material degradation and, consequently, the longevity of solar cells under practical use and storage conditions.[152] The vast volume of data generated by automated experimental technology offers a comprehensive insight into the intricate dynamics of OSCs. This extensive dataset is invaluable, capturing the nuanced behavior of OSCs under various conditions. Such detailed information is crucial for constructing robust predictive models that can accurately predict OSCs' responses to different stimuli. The AMANDA L1, developed by Du et al.[84–86] serves as an excellent example of evaluating OSC complex performance. We know that the role of AI in scientific research is growing, with AI now capable of automatically detecting experimental variables potentially missed by human researchers. A notable example is the Autonomous Laboratory (A-Lab), an AI-directed facility that uses robotics to synthesize new materials.[153] The A-Lab operates with minimal human supervision, enabling rapid discovery of new materials. It harnesses computation, historical data from literature, ML, and active learning to analyze experiments conducted by robots. Meanwhile, the convergence of automated experimental technology with AI in the research and development of OSCs represents a potent synergy that is poised to expedite progress in this domain. This combination of ML approaches and automated experimental technology is anticipated to emerge as a significant research avenue in the future.

6. Summary

In conclusion, the integration of computational methods with ML is set to revolutionize materials science, particularly in the domain of OSCs. ML as a preferred approach for forecasting material performance promises substantial benefits in steering experimental research, hastening the identification of novel materials, and deepening our comprehension of material characteristics and behaviors. In this review, we have outlined the current challenges and proposed corresponding research directions for the future: developing extended ML models based on RF model, generating a specific, small-scale experimental dataset, integrating multiple types of descriptors, and merging ML methods with automated experimental techniques. We hope that these potential research directions can benefit the prediction of OSC devices using ML approaches, significantly impacting various industries that rely on advanced materials.

Acknowledgements

The authors acknowledge the support of the Science and Technology Research Program of the Chongqing Education Commission of China (grant nos. KJZD-K202201403, KJQN202401431), the Natural Science Foundation of Chongqing, China (grant nos. CSTB2022NSCQ-MSX0395, CSTB2024NSCQ-MSX0850), and the Postgraduate Research & Practice Innovation Program of Yancheng Institute of Technology (KYCX24_XY003).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

machine learning, organic solar cells, power conversion efficiencies, predictive models

Received: August 2, 2024
Revised: September 4, 2024
Published online: October 9, 2024

[1] J. Fu, Q. Yang, P. Huang, S. Chung, K. Cho, Z. Kan, H. Liu, X. Lu, Y. Lang, H. Lai, F. He, P. W. K. Fong, S. Lu, Y. Yang, Z. Xiao, G. Li, *Nat. Commun.* **2024**, *15*, 1830.
[2] L. Ma, S. Zhang, J. Zhu, Z. Chen, T. Zhang, X. Hao, J. Hou, *Joule* **2024**, *8*, 1.
[3] Y. Z. Yang, C. Yao, L. Li, M. L. Bo, J. F. Zhang, C. Peng, J. S. Wang, *RSC Adv.* **2020**, *10*, 12004.
[4] Y. Liu, J. Zhao, Z. Li, C. Mu, W. Ma, H. Hu, K. Jiang, H. Lin, H. Ade, H. Yan, *Nat. Commun.* **2014**, *5*, 5293.
[5] M. Deng, X. Xu, Y. Duan, W. Qiu, L. Yu, R. Li, Q. Peng, *Adv. Mater.* **2024**, *36*, 2308216.
[6] B. Zou, H. M. Ng, H. Yu, P. Ding, J. Yao, D. Chen, S. H. Pun, H. Hu, K. Ding, R. Ma, M. Qammar, W. Liu, W. Wu, J. Y. L. Lai, C. Zhao, M. Pan, L. Guo, J. E. Halpert, H. Ade, G. Li, H. Yan, *Adv. Mater.* **2024**, *36*, 2405404.
[7] H. Jinno, K. Fukuda, X. Xu, S. Park, Y. Suzuki, M. Koizumi, T. Yokota, I. Osaka, K. Takimiya, T. Someya, *Nat. Energy* **2017**, *2*, 780.
[8] C. Zhang, X. Zhong, X. Sun, J. Lv, Y. Ji, J. Fu, C. Zhao, Y. Yao, G. Zhang, W. Deng, K. Wang, G. Li, H. Hu, *Adv. Sci.* **2024**, *11*, 2401313.
[9] A. These, L. J. A. Koster, C. J. Brabec, V. M. Le Corre, *Adv. Energy Mater.* **2024**, *14*, 2400055.
[10] G. Cai, Y. Li, Y. Fu, H. Yang, L. Mei, Z. Nie, T. Li, H. Liu, Y. Ke, X. L. Wang, J. L. Bredas, M. C. Tang, X. Chen, X. Zhan, X. Lu, *Nat. Commun.* **2024**, *15*, 2784.

[11] E.-Y. Shin, H. J. Son, *Nat. Energy* **2024**, *9*, 767.
[12] J. Cheng, C. Guo, L. Wang, Y. Fu, D. Li, C. Chen, Z. Gan, Y. Sun, D. Liu, W. Li, T. Wang, *Joule* **2024**, *8*, 2250.
[13] L. Xu, S. Li, W. Zhao, Y. Xiong, J. Yu, J. Qin, G. Wang, R. Zhang, T. Zhang, Z. Mu, J. Zhao, Y. Zhang, S. Zhang, V. Kuvondikov, E. Zakhidov, Q. Peng, N. Wang, G. Xing, F. Gao, J. Hou, W. Huang, J. Wang, *Adv. Mater.* **2024**, *36*, 2403476.
[14] Y. Yang, C. Yao, L. Li, M. Bo, J. Zhang, C. Peng, Z. Huang, J. Wang, *Dyes Pigm.* **2020**, *181*, 108542.
[15] C. Yao, X. Xu, J. Wang, L. Shi, L. Li, *ACS Appl. Mater. Interfaces* **2013**, *5*, 1100.
[16] X. Jiang, X. Wang, Y. Wang, G. Ran, W. Liu, H. Lu, H. Li, N. Wei, Z. Wei, Y. Lin, Z. Ma, Y. Liu, W. Zhang, X. Xu, Z. Bo, *Adv. Funct. Mater.* **2024**, *34*, 2406744.
[17] H. Hu, S. Liu, J. Xu, R. Ma, Z. Peng, T. A. D. Pena, Y. Cui, W. Liang, X. Zhou, S. Luo, H. Yu, M. Li, J. Wu, S. Chen, G. Li, Y. Chen, *Angew. Chem., Int. Ed. Engl.* **2024**, *63*, e202400086.
[18] F. L. Cheng, Y. J. Cui, F. Ding, Z. Chen, Q. Xie, X. X. Xia, P. P. Zhu, X. H. Lu, H. M. Zhu, X. F. Liao, Y. W. Chen, *Adv. Mater.* **2023**, *35*, 2300820.
[19] D. H. Li, N. Deng, Y. W. Fu, C. H. Guo, B. J. Zhou, L. Wang, J. Zhou, D. Liu, W. Li, K. Wang, Y. M. Sun, T. Wang, *Adv. Mater.* **2023**, *35*, 2208211.
[20] D. G. Gulevich, I. R. Nabiev, P. S. Samokhvalov, *Mater. Today Chem.* **2024**, *35*, 101837.
[21] G. Han, Y. Yi, *Angew. Chem., Int. Ed.* **2022**, *61*, e202213953.
[22] M. Chen, Z. Yin, Z. Shan, X. Zheng, L. Liu, Z. Dai, J. Zhang, S. Liu, Z. Xu, *J. Energy Chem.* **2024**, *94*, 254.
[23] S. Escalera, O. Pujol, P. Radeva, *J. Mach. Learn. Res.* **2010**, *11*, 661.
[24] M. Zhang, Z. Zhou, *IEEE Trans. Knowl. Date En.* **2014**, *26*, 1819.
[25] E. M. Williamson, R. L. Brutchey, *Inorg. Chem.* **2023**, *62*, 16251.
[26] E. J. Braham, J. Cho, K. M. Forlano, D. F. Watson, R. Arròyave, S. Banerjee, *Chem. Mater.* **2019**, *31*, 3281.
[27] M. Golmohammadi, M. Aryanpour, *Mater. Today Commun.* **2023**, *35*, 105494.
[28] J. C. Hummelen, C. J. Brabec, A. C. Cravino, D. Meissner, L. Sanchez, *Adv. Funct. Mater.* **2001**, *11*, 374.
[29] J. Liu, Y. Shi, Y. Yang, *Adv. Funct. Mater.* **2001**, *11*, 420.
[30] A. Gadisa, M. Svensson, M. R. Andersson, O. Inganäs, *Appl. Phys. Lett.* **2004**, *84*, 1609.
[31] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, *Adv. Mater.* **2006**, *18*, 789.
[32] F. H. Alharbi, S. N. Rashkeev, F. El-Mellouhi, H. P. Lüthi, N. Tabet, S. Kais, *npj Comput. Mater.* **2015**, *1*, 15003.
[33] N. S. Sariciftci, L. Smilowitz, A. J. Heeger, F. Wudl, *Science* **1992**, *258*, 1474.
[34] K. Sun, Z. Xiao, S. Lu, W. Zajaczkowski, W. Pisula, E. Hanssen, J. M. White, R. M. Williamson, J. Subbiah, J. Ouyang, A. B. Holmes, W. W. Wong, D. J. Jones, *Nat. Commun.* **2015**, *6*, 6013.
[35] M. Zhang, X. Guo, W. Ma, H. Ade, J. Hou, *Adv. Mater.* **2015**, *27*, 4655.
[36] Y. Lin, J. Wang, Z. G. Zhang, H. Bai, Y. Li, D. Zhu, X. Zhan, *Adv. Mater.* **2015**, *27*, 1170.
[37] J. Yuan, T. Huang, P. Cheng, Y. Zou, H. Zhang, J. L. Yang, S. Y. Chang, Z. Zhang, W. Huang, R. Wang, D. Meng, F. Gao, Y. Yang, *Nat. Commun.* **2019**, *10*, 570.
[38] C. Yao, Y. Yang, L. Li, M. Bo, C. Peng, J. Wang, *J. Mater. Chem. A* **2019**, *7*, 18150.
[39] J. Yuan, Y. Zhang, L. Zhou, G. Zhang, H.-L. Yip, T.-K. Lau, X. Lu, C. Zhu, H. Peng, P. A. Johnson, M. Leclerc, Y. Cao, J. Ulanski, Y. Li, Y. Zou, *Joule* **2019**, *3*, 1140.
[40] J. Fu, P. W. K. Fong, H. Liu, C. S. Huang, X. Lu, S. Lu, M. Abdelsamie, T. Kodalle, C. M. Sutter-Fella, Y. Yang, G. Li, *Nat. Commun.* **2023**, *14*, 1760.
[41] D. Luo, Z. Jiang, W. L. Tan, L. Zhang, L. Li, C. Shan, C. R. McNeill, P. Sonar, B. Xu, A. K. K. Kyaw, *Adv. Energy Mater.* **2023**, *13*, 2203402.
[42] J. Yi, G. Zhang, H. Yu, H. Yan, *Nat. Rev. Mater.* **2023**, *9*, 46.
[43] L. Ma, S. Zhang, J. Zhu, J. Wang, J. Ren, J. Zhang, J. Hou, *Nat. Commun.* **2021**, *12*, 5093.
[44] N. Yang, Y. Cui, Y. Xiao, Z. Chen, T. Zhang, Y. Yu, J. Ren, W. Wang, L. Ma, J. Hou, *Angew. Chem., Int. Ed. Engl.* **2024**, *63*, e202403753.
[45] N. Yang, Y. Cui, T. Zhang, C. An, Z. Chen, Y. Xiao, Y. Yu, Y. Wang, X. T. Hao, J. Hou, *J. Am. Chem. Soc.* **2024**, *146*, 9205.
[46] J. Wang, P. Xue, Y. Jiang, Y. Huo, X. Zhan, *Nat. Rev. Chem.* **2022**, *6*, 614.
[47] X. Gu, X. Zhang, H. Huang, *Angew. Chem.* **2023**, *62*, e202308496.
[48] H. Chen, S. Y. Jeong, J. Tian, Y. Zhang, D. R. Naphade, M. Alsufyani, W. Zhang, S. Griggs, H. Hu, S. Barlow, H. Y. Woo, S. R. Marder, T. D. Anthopoulos, I. McCulloch, Y. Lin, *Energy Environ. Sci.* **2023**, *16*, 1062.
[49] Y. Imamura, M. Suganuma, M. Hada, *J. Phys. Chem. C* **2019**, *123*, 17678.
[50] B. L. Greenstein, G. R. Hutchison, *J. Phys. Chem. Lett.* **2022**, *13*, 4235.
[51] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt, A. Aspuru-Guzik, *Energy Environ. Sci.* **2011**, *4*, 4849.
[52] J. Perea, S. Langner, M. Salvador, B. Sanchez, N. Li, C. Zhang, G. Jarvás, J. Kontos, A. Dallos, A. Aspuru-Guzik, C. Brabec, *J. Phys. Chem. C* **2017**, *121*, 18153.
[53] H. Sahu, W. Rao, A. Troisi, H. Ma, *Adv. Energy Mater.* **2018**, *8*, 1801032.
[54] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.* **2018**, *9*, 2639.
[55] 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.
[56] T. Wang, R. Sun, M. Shi, F. Pan, Z. Hu, F. Huang, Y. Li, J. Min, *Adv. Energy Mater.* **2020**, *10*, 10.
[57] S. P. Peng, Y. Zhao, *J. Chem. Inf. Model* **2019**, *59*, 4993.
[58] Y. Wu, J. Guo, R. Sun, J. Min, *npj Comput. Mater.* **2020**, *6*, 120.
[59] P. Malhotra, S. Biswas, F.-C. Chen, G. D. Sharma, *Sol. Energy* **2021**, *228*, 175.
[60] A. Mahmood, A. Irfan, J.-L. Wang, *J. Mater. Chem. A* **2022**, *10*, 4170.
[61] X. Liu, Y. Shao, T. Lu, D. Chang, M. Li, W. Lu, *Mater. Des.* **2022**, *216*, 110561.
[62] R. Suthar, T. Abhijith, S. Karak, *J. Mater. Chem. A* **2023**, *11*, 22248.
[63] C.-R. Zhang, M. Li, M. Zhao, J.-J. Gong, X.-M. Liu, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *J. Appl. Phys.* **2023**, *134*, 153104.
[64] J.-H. Li, C.-R. Zhang, M.-L. Zhang, X.-M. Liu, J.-J. Gong, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *Org. Electron.* **2024**, *125*, 106988.
[65] T. Mubashir, M. H. Tahir, Z. Shafiq, A. Z. Dewidar, H. O. El-ansary, M. R. S. A. Janjua, *J. Photoch. Photobio. A* **2024**, *447*, 115285.
[66] R. Cao, C.-R. Zhang, M. Li, X.-M. Liu, M.-L. Zhang, J.-J. Gong, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *Sol. RRL.* **2024**, *8*, 2400370.
[67] S. Huang, Q. Li, S. Li, C. Li, H. Tan, Y. Xie, *Chem. Commun.* **2024**, *60*, 4521.
[68] J. A. Bhutto, B. Siddique, I. M. Moussa, M. A. El-Sheikh, Z. Hu, G. Yurong, *Heliyon* **2024**, *10*, e30473.
[69] C. Yao, C. Peng, Y. Yang, L. Li, M. Bo, J. Wang, *J. Phys. Chem. C* **2018**, *122*, 22273.
[70] C. Xu, C. Yao, S. Zheng, *J. Mater. Chem. C* **2021**, *9*, 14637.
[71] M. K. Gish, C. D. Karunasena, J. M. Carr, W. P. Kopcha, A. L. Greenaway, A. A. Mohapatra, J. Zhang, A. Basu, V. Brosius, S. M. Pratik, J. L. Bredas, V. Coropceanu, S. Barlow, S. R. Marder, A. J. Ferguson, O. G. Reid, *J. Phys. Chem. C* **2024**, *128*, 6392.
[72] Y. Jiang, S. Sun, R. Xu, F. Liu, X. Miao, G. Ran, K. Liu, Y. Yi, W. Zhang, X. Zhu, *Nat. Energy* **2024**, *9*, 975.

[73] J. Wu, W. Ma, T. Li, J. Yan, Z. He, Y. Cao, *ACS Appl. Mater. Interfaces* **2024**, *16*, 29466.
[74] A. Wieczorek, Y. Liu, H. H. Cho, K. Sivula, *J. Phys. Chem. Lett.* **2024**, *15*, 6347.
[75] W. Song, Q. Ye, Z. Chen, J. Ge, L. Xie, Z. Ge, *Adv. Mater.* **2024**, *36*, 2311170.
[76] R. Noruzi, S. Ghadai, O. R. Bingol, A. Krishnamurthy, B. Ganapathysubramanian, *Comput. Aided Des.* **2020**, *118*, 102771.
[77] S. G. Jung, G. Jung, J. M. Cole, *J. Chem. Phys.* **2023**, *159*, 194106.
[78] H. Wang, J. Feng, Z. Dong, L. Jin, M. Li, J. Yuan, Y. Li, *npj Comput. Mater.* **2023**, *9*, 200.
[79] S. Zhong, W. Hsu, H. Chen, T. Yang, J. Yi, C. Zhu, S. Yin, Z. Li, L. Gao, J. Lin, L. Ying, N. Li, *Sol. RRL* **2024**, *8*, 2400288.
[80] J. Wang, J. R. R. A. Martins, X. Du, *Aerosp. Sci. Technol.* **2024**, *150*, 109214.
[81] M. Vubangsi, A. S. Mubarak, F. Al-Turjman, *Energy Rep.* **2024**, *11*, 3824.
[82] S. Zhang, S. Wei, Z. Liu, T. Li, C. Li, X. L. Huang, C. Wang, Z. Xie, O. A. Al-Hartomy, A. A. Al-Ghamdi, S. Wageh, J. Gao, Y. Tang, H. Wang, Q. Wang, H. Zhang, *Mater. Today Phys.* **2022**, *27*, 100812.
[83] H. Eldeeb, M. Maher, R. Elshawi, S. Sakr, *Expert Syst. Appl.* **2024**, *243*, 122877.
[84] X. Du, L. Lüer, T. Heumueller, J. Wagner, C. Berger, T. Osterrieder, J. Wortmann, S. Langner, U. Vongsaysy, M. Bertrand, N. Li, T. Stubhan, J. Hauch, C. J. Brabec, *Joule* **2021**, *5*, 495.
[85] J. Wagner, C. G. Berger, X. Du, T. Stubhan, J. A. Hauch, C. J. Brabec, *J. Mater. Sci.* **2021**, *56*, 16422.
[86] X. Du, L. Lüer, T. Heumueller, A. Classen, C. Liu, C. Berger, J. Wagner, V. M. Le Corre, J. Cao, Z. Xiao, L. Ding, K. Forberich, N. Li, J. Hauch, C. J. Brabec, *InfoMat.* **2024**, *6*, 12554.
[87] N. M. O'Boyle, C. M. Campbell, G. R. Hutchison, *J. Phys. Chem. C* **2011**, *115*, 16200.
[88] F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang, J. Aires-De-Sousa, *J. Chem. Inf. Model* **2016**, *57*, 11.
[89] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241.
[90] J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Román-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao, A. Aspuru-Guzik, *Energy Environ. Sci.* **2014**, *7*, 698.
[91] W. Sun, M. Li, Y. Li, Z. Wu, Y. Sun, S. Lu, Z. Xiao, B. Zhao, K. Sun, *Adv. Theory Simul.* **2018**, *2*, 1800116.
[92] S. A. Lopez, B. Sanchez-Lengeling, J. D. G. Soares, A. Aspuru-Guzik, *Joule* **2017**, *1*, 856.
[93] P. C. St John, C. Phillips, T. W. Kemper, A. N. Wilson, Y. Guan, M. F. Crowley, M. R. Nimlos, R. E. Larsen, *J. Chem. Phys.* **2019**, *150*, 234111.
[94] R. J. Richards, A. Paul, *Sol. Energy* **2021**, *224*, 43.
[95] J.-M. Liao, H.-H. G. Tsai, *Sol. RRL* **2024**, *8*, 2400287.
[96] Y. Miyake, A. Saeki, *J. Phys. Chem. Lett.* **2021**, *12*, 12391.
[97] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik, *Sci. Data* **2016**, *3*, 160086.
[98] H. Sahu, H. Ma, *J. Phys. Chem. Lett.* **2019**, *10*, 7277.
[99] E. A. J. Abadi, H. Sahu, S. M. Javadpour, M. Goharimanesh, *Mater. Today Energy* **2022**, *25*, 100969.
[100] M. H. Lee, *Adv. Energy Mater.* **2019**, *9*, 1900891.
[101] Q. Zhao, Y. Shan, H. Zhou, G. Zhang, W. Liu, *Sol. Energy* **2023**, *265*, 112115.
[102] M. Li, C.-R. Zhang, M.-L. Zhang, J.-J. Gong, X.-M. Liu, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *Phys. Status Solidi A* **2024**, *221*, 2400008.
[103] A. A. A. Torimtubun, M. J. Alonso-Navarro, A. Quesada-Ramírez, X. Rodríguez-Martínez, J. L. Segura, A. R. Goñi, M. Campoy-Quiles, *Sol. RRL* **2024**, *8*, 2400213.
[104] A. Mahmood, Y. Sandali, J.-L. Wang, *PCCP* **2023**, *25*, 10417.
[105] Y. Fu, T. H. Lee, Y.-C. Chin, R. A. Pacalaj, C. Labanti, S. Y. Park, Y. Dong, H. W. Cho, J. Y. Kim, D. Minami, J. R. Durrant, J.-S. Kim, *Nat. Commun.* **2023**, *14*, 1870.
[106] F. Ahmad, A. Mahmood, I. H. El Azab, N. Ahmad, M. H. H. Mahmoud, Z. M. El-Bahy, *J. Photochem. Photobiol. A* **2024**, *453*, 115670.
[107] Y. Iwasaki, I. Takeuchi, V. Stanev, A. G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya, K.-I. Uchida, E. Saitoh, S. Yorozu, *Sci. Rep.* **2019**, *9*, 2751.
[108] S. L. Benjamin, A.-G. Alán, *Science* **2018**, *361*, 360.
[109] A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori, M. S. Rao, *Chem. Res. Toxicol.* **2020**, *33*, 20.
[110] M. Saqib, M. Rani, T. Mubashir, M. H. Tahir, M. Maryam, A. Mushtaq, R. Razzaq, M. A. El-Sheikh, H. O. Elansary, *Opt. Mater.* **2024**, *150*, 115295.
[111] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput. Aided Mol. Des.* **2016**, *30*, 595.
[112] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513.
[113] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, arXiv: 1704.01212. **2017**, 1263.
[114] A. Paul, D. Jha, R. Al-Bahrani, W. K. Liao, A. Choudhary, A. Agrawal, arXiv: 1903.03178. **2019**.
[115] P. B. Jorgensen, M. N. Schmidt, O. Winther, *Mol. Inform.* **2018**, *37*, 1700133.
[116] Z. Alperstein, A. Cherkasov, J. T. Rolfe, arXiv: 1905.13343. **2019**.
[117] K. Mohammedsaleh Katubi, M. Saqib, M. Sulaman, Z. A. Alrowaili, M. S. Al-Buriahi, *Chem. Phys.* **2024**, *582*, 112295.
[118] M. Liu, X. Han, H. Chen, Q. Peng, H. Huang, *Nat. Commun.* **2023**, *14*, 2500.
[119] D. Padula, A. Troisi, *Adv. Energy Mater.* **2019**, *9*, 1902463.
[120] E. O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, *Adv. Funct. Mater.* **2015**, *25*, 6495.
[121] W. Sun, Y. Zheng, Q. Zhang, K. Yang, H. Chen, Y. Cho, J. Fu, O. Odunmbaku, A. A. Shah, Z. Xiao, *J. Phys. Chem. Lett.* **2021**, *12*, 8847.
[122] F. C. Chen, *Int. J. Polym. Sci.* **2019**, *2019*, 1.
[123] F. M. A. Alzahrani, S. Naeem, N. Khan, B. Siddique, M. Faizan Nazar, T. Kadyrov, Z. A. Alrowaili, M. S. Al-Buriahi, *Comp. Mater. Sci.* **2024**, *239*, 112984.
[124] E. O. Pyzer-Knapp, G. N. Simm, A. Aspuru-Guzik, *Mater. Horiz.* **2016**, *3*, 226.
[125] S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, A. Aspuru-Guzik, *Joule* **2017**, *1*, 857.
[126] M. H. Lee, *Org. Electron.* **2020**, *76*, 105465.
[127] M. H. Lee, *Adv. Intell. Syst.* **2020**, *2*, 1900108.
[128] M. H. Lee, *Energy Technol.* **2020**, *8*, 1900974.
[129] Z.-W. Zhao, M. del Cueto, Y. Geng, A. Troisi, *Chem. Mater.* **2020**, *32*, 7777.
[130] D. Huang, K. Wang, Z. Li, H. Zhou, X. Zhao, X. Peng, J. Wu, J. Liang, J. Meng, L. Zhao, *Chem. Eng. J.* **2023**, *475*, 145958.
[131] X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov, F. Cheng, *Nat. Mach. Intell.* **2022**, *4*, 1004.
[132] L. Rampáek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, D. Beaini, arXiv.2205.12454 **2022**.
[133] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, K. Sun, *Sci. Adv.* **2019**, *5*, eaay4275.

[134] M. Seifrid, S. Lo, D. G. Choi, G. Tom, M. L. Le, K. Li, R. Sankar, H.-T. Vuong, H. Wakidi, A. Yi, Z. Zhu, N. Schopp, A. Peng, B. R. Luginbuhl, T.-Q. Nguyen, A. Aspuru-Guzik, *J. Mater. Chem. A* **2024**, *12*, 14540.
[135] Z. Hui, M. Wang, X. Yin, Y. N. Wang, Y. L. Yue, *Comp. Mater. Sci.* **2023**, *226*, 112215.
[136] G. Cheng, X.-G. Gong, W.-J. Yin, *Nat. Commun.* **2022**, *13*, 1492.
[137] B. L. Greenstein, G. R. Hutchison, *J. Phys. Chem. C* **2023**, *127*, 6179.
[138] J. Sun, D. Li, Y. Wang, T. Xie, Y. Zou, H. Lu, Z. Zhang, *J. Mater. Chem. A* **2024**, *12*, 21813.
[139] D. Padula, J. D. Simpson, A. Troisi, *Mater. Horiz.* **2019**, *6*, 343.
[140] H. Chen, Z. Zhang, P. Wang, Y. Zhang, K. Ma, Y. Lin, T. Duan, T. He, Z. Ma, G. Long, C. Li, B. Kan, Z. Yao, X. Wan, Y. Chen, *Energy Environ. Sci.* **2023**, *16*, 1773.
[141] A. Mahmood, J.-L. Wang, *Energy Environ. Sci.* **2021**, *14*, 90.
[142] K. Mohammedsaleh Katubi, A. Muhammad Rouf, B. Siddique, M. Faizan Nazar, G. Jillani Ansari, Z. A. Alrowaili, M. S. Al-Buriahi, *Comput. Mater. Sci.* **2024**, *241*, 113037.
[143] Y. Zhang, Y. Lang, G. Li, *EcoMat.* **2023**, *5*, e12281.
[144] X. Zhao, M. Lei, K. Wang, X. Peng, Z. Li, H. Zhou, Z. Peng, Z. Chen, J. Deng, K. Zhang, D. Huang, J. Liang, *AIP Adv.* **2024**, *14*, 065325.
[145] J. Zhang, L. He, Y. Xiong, S. Huang, B. Xu, S. Ma, X. Xiang, H. Fu, J. Kai, Z. Wu, S. Zhao, *npj Comput. Mater.* **2024**, *10*, 162.
[146] C. Yao, X. Li, Y. Z. Yang, L. Li, M. L. Bo, C. Peng, J. S. Wang, *J. Mater. Chem. A* **2022**, *10*, 15999.
[147] C. Yao, Y. Yang, Y. Ou, J. Wang, *J. Phys. Chem. C* **2023**, *127*, 24039.
[148] K. Kranthiraja, A. Saeki, *Adv. Funct. Mater.* **2021**, *31*, 2011168.
[149] Y. Wen, Y. Liu, B. Yan, T. Gaudin, J. Ma, H. Ma, *J. Phys. Chem. Lett.* **2021**, *12*, 4980.
[150] H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang, H. Ma, *J. Mater. Chem. A* **2019**, *7*, 17480.
[151] P. Malhotra, K. Khandelwal, S. Biswas, F.-C. Chen, G. D. Sharma, *J. Mater. Chem. C* **2022**, *10*, 17781.
[152] C. Yan, J. Qin, Y. Wang, G. Li, P. Cheng, *Adv. Energy Mater.* **2022**, *12*, 2201087.
[153] N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, G. Ceder, *Nature* **2023**, *624*, 86.
[154] A. Chen, X. Zhang, Z. Zhou, *InfoMat.* **2020**, *2*, 553.
[155] S. G. Jung, G. Jung, J. M. Cole, *J. Chem. Phys.* **2023**, *159*, 194106.
[156] F. C. Chen, *Int. J. Polym. Sci.* **2019**, *2019*, 4538514.
[157] A. Paul, D. Jha, R. Al-Bahrani, W. K. Liao, A. Choudhary, A. Agrawal, in *2019 Inter. Joint Conf. on Neural Networks (IJCNN)* , Budapest, Hungary, 14-19 July **2019**, pp. 1–8.

Yang Jiang is now a master's student in the School of Materials Science and Engineering at Yancheng Institute of Technology, under the guidance of Professor Chuang Yao and Associate Professor Jinshan Wang. He primarily investigates organic solar cells (OSCs). His research interests include using machine learning approaches to predict the power conversion efficiency of OSCs and designing quad-rotor-shaped nonfullerene acceptors.

Chuang Yao is a professor in the School of Materials Science and Engineering at Yangtze Normal University (YZNU), China. He received his Ph.D. degree at the University of Science and Technology Beijing, China, in 2016. After that, he joined YZNU as an associate professor and was promoted to full professor in 2022. His research interests focus on developing high-efficiency organic optoelectronic materials, especially for employing theoretical calculations and machine learning techniques to facilitate the development of photoactive materials.

Yezi Yang works as a laboratory technician in the School of Materials Science and Engineering at Yangtze Normal University (YZNU), China. She received her master's degree from Xiangtan University in 2014. Her research interests focused on designing newly high-performance nonfullerene electron acceptor materials for organic solar cells.

Jinshan Wang is an Associate Professor of Materials Science and Engineering at Yancheng Institute of Technology. He obtained his doctoral degree from Beijing University of Science and Technology in 2015.Then, he joined Yancheng Institute of Technology through the "High level Innovative Talent Introduction Project." He is currently focused on molecular design and synthesis of new optoelectronic functional materials for optoelectronic devices, such as organic light-emitting diodes and organic solar cells.