Journal of
Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d2tc03276g

Opportunities and challenges for machine learning to select combination of donor and acceptor materials for efficient organic solar cells†

Prateek Malhotra,[a] Kanupriya Khandelwal,[a] Subhayan Biswas,[a] Fang-Chung Chen[bc] and Ganesh D. Sharma *[ad]

Received 5th August 2022,
Accepted 19th October 2022

DOI: 10.1039/d2tc03276g

rsc.li/materials-c

Organic solar cells (OSCs) have witnessed significant improvement in power conversion efficiency (PCE) in the last decade. The structural flexibility of organic semiconductors provides vast search space for potential candidates of OSCs, but discovering new materials from search space with traditional approaches such as DFT is computationally expensive and time-consuming. Machine learning (ML) is extensively used in OSCs to accelerate productivity and materials discovery. ML is gaining more attention due to the availability of large datasets, improved algorithms, and exponentially growing computational power. In this review, current progress, opportunity, and challenges for ML in OSCs have been identified. Given the rapid advances in this field, impactful techniques that have been useful in extracting meaningful insights are discussed. Finally, we elaborate upon the bottlenecks of the ML workflow with respect to data size, model interpretability, and extrapolation.

1. Introduction

Organic solar cells (OSCs) have come a long way in their quest to become a viable alternative to more expensive conventional solar cell technology.[1,2] Being flexible, lightweight, eco-friendly, and semi-transparent, OSCs hold promising potential for various applications.[3–8] With the advent of non-fullerene small molecule acceptors (NFSMAs),[9–13] especially Y-series NFSMAs,[14–16] bulk heterojunction (BHJ) based OSCs have already reached power conversion efficiencies (PCEs) in the range of 18–19% for single junction binary and ternary systems[17–23] and are exceeding 20% for tandem configuration.[24] Sharp absorption

[a] Department of Physics, The LNM Institute of Information Technology, Jamdoli, Jaipur, 302031, Rajasthan, India. E-mail: gdsharma273@gmail.com, gdsharma@lnmiit.ac.in

[b] Department of Photonics, National Yang Ming Chiao Tung University, Hsinchu, 30010, Taiwan

[c] Center for Emergent Functional Matter Science, National Yang Ming Chiao Tung University, Hsinchu, 30010, Taiwan

[d] Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jamdoli, Jaipur, 302031, Rajasthan, India

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2tc03276g

Prateek Malhotra

Prateek Malhotra is a PhD scholar in the Department of Physics at The LNM Institute of Information Technology, India, under the supervision of Prof. Ganesh D. Sharma and Dr Subhayan Biswas. His research focuses on developing high-efficiency organic solar cells and machine learning applications for the prediction of photovoltaic parameters of organic solar cells.

Kanupriya Khandelwal

Kanupriya Khandelwal is a PhD Scholar in the Department of Physics at The LNM Institute of Information Technology, Jaipur, India, under the supervision of Prof. G. D. Sharma and Dr Subhayan Biswas. Her main research interests are in the optoelectronic properties of organic materials for application in photovoltaic devices. She is currently working on semi-transparent organic solar cells and their BIPV (Building Integrated Photovoltaics) application.

onset, bandgap tunability, high absorption, and low energy losses are some of the benefits of NFSMAs over fullerene derivatives that have propelled them to the forefront of technological advancement and potential commercialization of OSCs at low cost.[25–28]

The most efficient OSCs are based on the concept of BHJ, in which the active layer is sandwiched between the anode and the cathode. The hole transport layer (HTL) is inserted between the BHJ layer and the anode, and the electron transport layer (ETL) is inserted between the BHJ layer and the cathode to enhance the charge collection. The BHJ active layer is a mix of donor (p-type organic semiconductor) and acceptor (n-type organic semiconductor), which provides an appropriate phase separation for efficient exciton dissociation into free charge carriers and interpenetrating network pathways for charge transport towards respective electrodes. In general, the energy conversion in OSCs based on the bulk heterojunction active layer is accomplished by the following main consecutive steps: (i) absorption of photons by the BHJ active layer and exciton generation, (ii) exciton diffusion to D/A interfaces, (iii) exciton dissociation, (iv) transport of electrons and holes towards the cathode and anode, and (v) charge extraction. The step (vi) charge recombination, would lead to a decrease in device efficiency; these processes are presented in Fig. 1. Considering

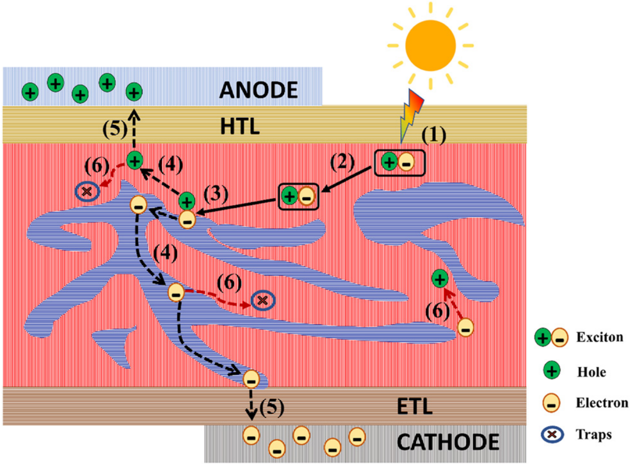

Fig. 1 Photocurrent generation in BHJ-OSCs. (1) Exciton generation, (2) exciton diffusion, (3) exciton dissociation, (4) charge transportation, (5) charge extraction, and (6) charge recombination.

Subhayan Biswas

Dr Subhayan Biswas is an associate professor in the Department of Physics at The LNM Institute of Information Technology, Jaipur, India. After completing his PhD from the Indian Association for the Cultivation of Science, Kolkata, India, he did post-doctoral research in Inha University (South Korea), National Sun Yat-Sen University & National Taiwan University (Taiwan) and Toyama University (Japan). His research field is nanostructured solar cells including dye-sensitized solar cells, quantum-dot solar cells and organic solar cells.

Fang-Chung Chen

Prof. Fang-Chung Chen is currently a Professor and Chairman of Department of Photonics (DoP), National Yang Ming Chiao Tung University (NYCU). He joined DoP, National Chiao Tung University (NCTU, now NYCU) in February 2004. He was also the chairman of Degree Program of Flat Panel Display Technology, NCTU. He has published more than 130 journal papers, 100 conference papers, and 5 book chapters, and owns 14 patents. Prof. Chen is a Fellow of the Royal Society of Chemistry (FRSC). He is also the recipient of the 2021 IoT Innovation Award (Pen Wen Yuan Foundation), 2020 Y. Z. Hsu Scientific Paper Award and Award for Junior Research Investigators of Academia Sinica 2008, which is one of the most important awards for junior research investigators in all research fields in Taiwan. His h-index is 48 (Google Scholar). His research interests include flexible solar cells, organic electronics and materials, perovskite electronics, plasmonic materials and low-dimensional nanomaterials.

Ganesh D. Sharma

Prof. Ganesh D. Sharma has been working as a Senior Professor of Physics, Electronics Communication and Engineering and Dean (Research and development) at The LNM Institute of Information Technology, Jaipur (Raj.), India, since December 2014. Prof. Sharma obtained his PhD degree from Indian Institute of Technology in 1985 and after that he joined JNV University, Jodhpur (Raj), India, as Assistant Professor and subsequently as Professor. His area of research is organic solar cells, nanocrystalline organic–inorganic hybrid solar cells and fuel cells, and organic nanomaterials for energy conversion. He has published more than 330 research papers. He had research collaboration with different international and national research organizations.

all the processes, the PCE of the OSC depends on many properties of the organic semiconductors such as the optical absorption profile and molar extinction coefficient, charge carrier mobility, frontier energy levels of donor and acceptor, and electron–hole (exciton) binding energy (E_{bind}) and it is very important to construct more precise models using all appropriate and easily available descriptors.

Although OSCs have attained PCEs in the range of 19–20%, one faces new challenges in exploring numerous possible material combinations together with weight ratios between donor and acceptor and processing conditions. Once a market-competitive PCE is realized in the laboratory, the next step must be transferring this technology for manufacturing, which is challenging and time-consuming.[29,30] This demands a new research method that can rapidly explore enormous parameter space, ideally *via* industry-relevant methods. Many experimental methods, including screen printing, doctor blade, and slot-die, have been explored to cover the enormous composition space.[31–34]

Most efficient BHJ-OSCs employ the A–D–A structured NFSMAs as the acceptor and D–A copolymers as the donor. The structural diversity of these materials is of significant interest for materials scientists; however, it is not possible to completely survey the huge molecular space. Quantum chemical calculations and molecular dynamics simulations can provide an approximate evaluation of the optical and electrochemical properties of new materials;[35–37] however, the complex relationship between the structure and these properties of materials used for OSCs has hindered the efficient assessment of materials.

Machine learning (ML) is a subfield of rapidly developing artificial intelligence (AI) technology that seeks to create programs capable of learning from large data sets by employing various algorithms and statistical techniques. Such programs can then be used to do things like explore hidden patterns in data, build predictive models, and create guidelines for future research. In recent years, many publications on the application of ML techniques in materials science have been published, as these techniques have been widely used in material research to determine the properties and functions of existing materials or to discover new materials with more desirable functions. ML is getting a lot of attention these days because of its potential to boost output and aid in discovering new materials.[38] ML has been used extensively in fields including property prediction and material discovery in the field of OSCs.[39–43] The number of publications detailing the use of machine learning for the analysis or screening of data obtained experimentally or computationally has also expanded dramatically over the past couple of years. Since non-fullerene small molecule acceptors were discovered recently, earlier ML studies were performed entirely on polymer donors, and the acceptor was fullerene. The majority of datasets now being used for the study include polymer donors and small molecule non-fullerene acceptors. With an increase in data size, model performance is also improved.

Many high-performance active materials for OSCs are yet to be discovered, and material synthesis followed by device fabrication is expensive and time-consuming. We want to understand the functions that link material properties with structure. At the device fabrication stage, there are a number of tunable parameters such as the D : A ratio,[44] alignment between frontier molecular orbitals (FMOs),[45] miscibility,[46] film processing (spin-casting speed),[47] calcination temperature and duration,[48] active layer thickness[49], and processing additives.[47] Due to these tunable parameters, the study becomes more complex. To select the most optimal parameter combination, the one-variable-at-a-time method (Edisonian approach) is most often used.[40] In this trial-and-error approach, a lot of time and expensive materials are consumed without a guarantee of reaching the most optimal set of parameters. Such processes require a long time because they have to investigate the vast chemical space of materials. To counteract this, machine learning (ML) models have allowed researchers to investigate chemical space and its properties more effectively, saving time and money. The availability of large training datasets, superior algorithms, and ever-increasing processing power has excited the interest of materials scientists in ML. Photovoltaic parameters such as power conversion efficiency (PCE), short circuit current density (J_{SC}), open circuit voltage (V_{OC}), fill factor (FF), non-radiative voltage loss, and frontier molecular orbitals (FMO) have been used as target variables in many ML related studies for OSCs.

ML models are implemented for photovoltaic property prediction in OSCs using inputs such as molecular properties (MP), molecular descriptors (MD), fingerprints (FP), FMO, and molecular images. For the studies on OSCs, calculated descriptors/fingerprints of donor and acceptor are concatenated, and the total number of input descriptors for the ML model becomes very high and demands suitable feature engineering. Even after performing feature engineering, the remaining number of input descriptors are still high and cause degraded results by ML model. This is termed as ''curse of dimensionality''. Here comes the use of dimensionality reduction techniques such as principal component analysis (PCA), which is commonly used in the community. ML has granted the OSC research community entirely new abilities from predicting photovoltaic properties,[50] quantitative structure–property relationship (QSPR),[51,52] design of experiments (DOE),[47,53] novel polymer/NFA discovery,[54,55] and robotization of labs.[48,56] Most of the studies in the field are regression problems, and only a few are done for classification.

This review article is written to offer a synopsis of ML applications in OSCs. Section 2 describes the ML workflow, from data gathering to novel materials discovery. In addition, available datasets and their categories are also discussed. In Section 3, research papers on ML implementation in OSCs are analyzed and reviewed. The review is grouped on the basis of the acceptor type: (1) fullerene acceptor (FA), (2) non-fullerene acceptor (NFA), and (3) mix of FA and NFA. In Section 4, problems and future scope are discussed.

2. Machine learning workflow

To begin any ML project, professionals in the relevant field are consulted to help define the aims and targets of the models.

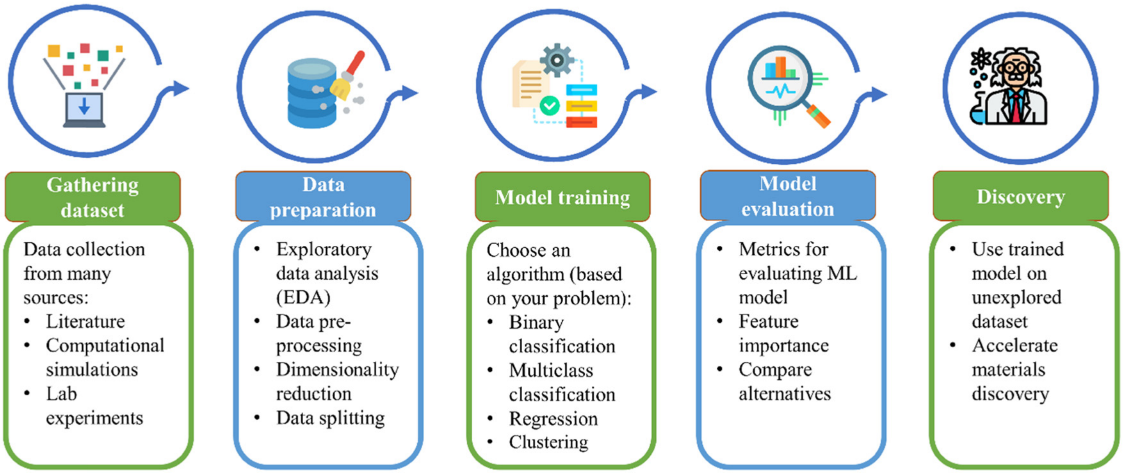
Gathering dataset
Data collection from many sources:
• Literature
• Computational simulations
• Lab experiments
Data preparation
• Exploratory data analysis (EDA)
• Data pre-processing
• Dimensionality reduction
• Data splitting
Model training
Choose an algorithm (based on your problem):
• Binary classification
• Multiclass classification
• Regression
• Clustering
Model evaluation
• Metrics for evaluating ML model
• Feature importance
• Compare alternatives
Discovery
• Use trained model on unexplored dataset
• Accelerate materials discovery

Fig. 2 Workflow of ML in OSCs.

This is the most critical phase since the target must be learnable from accessible information such as microscopic properties, molecular descriptors, molecular fingerprints, and DFT calculated descriptors. When the prediction objective is chosen incorrectly, it might lead to models with spuriously significant errors or models that aren't generalizable. After setting the research objectives, ML is implemented by systematically implementing the process summarized as (1) data gathering, (2) data preparation, (3) model training, (4) model evaluation, and (5) discovering and testing novel materials. Fig. 2 represents a high-level overview of the steps involved in the ML workflow.

2.1. Data gathering

For ML application in OSCs, input data are required to train the ML model, whether the task is prediction of photovoltaic properties or clustering of molecules into different categories. Early research in OSCs was dominated by fullerene acceptor-based solar cells because of high electron mobility due to their isotropic nature; thus, plenty of data could be easily collected from the literature for ML applications.

Prof. Alan Aspuru-Guzik and his team run the clean energy project. They created the Harvard clean energy project (HCEP) dataset consisting of 2.3 million compounds derived from 150 million DFT calculations to find high-efficiency OSC materials.[57–59] All these molecules are generated using 26 molecular building blocks. The properties include FMO, photovoltaic performance parameters (J_{sc}, V_{oc}, PCE), and stoichiometric formulas, and various studies have been performed using this dataset.[59,60] For high-throughput virtual screening, the HCEP dataset has been used by several research groups. Compounds in the HCEP dataset are considerably different and structurally simple from those found in the real world. Therefore, potential candidates for high-performance OSCs are less likely to be discovered from the HCEP dataset.

In 2006, the Scharber model[45] was developed to predict the PCE of fullerene based OSCs and the model has been used in various high-throughput screenings.[61–63] Since this model considers only a few electronic parameters and no other structural descriptors are considered, the performance of this model is relatively poor. Brabec *et al.*[64] proposed a modification to the Scharber model by taking the absorption of non-fullerene acceptors (NFAs) into account. In 2018, Ma *et al.*[50] created a FA based dataset of 280 OSCs with small molecule donors, and the same dataset was used by Goharimanesh *et al.*[65] in 2022 for interpretable ML models. Saeki *et al.*[66] in 2018 created a FA based dataset of 1200 OSCs with polymer donors for PCE prediction, and the same approach was used by Wei *et al.*[67] for 500 NFA based OSCs. In 2019, Chen[68] used the 1200 dataset by Saeki *et al.*[66] for virtual screening of conjugated polymers. In 2019, Troisi *et al.*[69] and Ma *et al.*[70] created FA datasets of 249 and 300 OSCs with small molecule donors. Ma *et al.*[71] used their earlier dataset (# = 300)[70] to unravel correlations between device performance parameters and molecular properties. In 2019, Lee[72] also developed a dataset of 124 fullerene derivative-based ternary OSCs for predicting PCE, and further in 2020, Lee[73] used the same dataset for V_{oc} prediction of ternary OSCs.

In 2015, Zhan *et al.* introduced efficient NFSMAs by synthesis of ITIC.[74] In 2017, Zhao *et al.*[75] presented the first NFA-based BHJ organic solar cell that outperformed the state-of-the-art solar cells with fullerenes as acceptors. Now NFA-based OSCs have crossed efficiency over 19%.[17,18] NFAs are used much more in recent studies due to their tunable absorption spectra, easily adjustable FMO, solubility, photostability, and thermal stability.[12,76–78] With the generation of new datasets for NFA-based OSCs, interest in NFA-based datasets is gathering momentum across the globe in ML applications.

Aspuru Guzik *et al.*[61] in 2017 created a dataset of 51 000 NFAs with PCDTBT as the donor material. In 2020, Min *et al.*[79] created a dataset of 565 polymer : NFA based OSCs for the discovery of new D : A combinations. In 2020, Lee[80] created a dataset of 135 NFA-based OSCs for PCE prediction, and later in 2022, Lee[81] used the same dataset (# = 135) to identify the correlation between FMO and open circuit voltage of OSCs. In 2021, Saeki *et al.*[82] created a dataset of 566 polymer donor : NFA based OSCs and later in the same year increased their dataset to 1318[83] for comparison. Later in 2022, Saeki *et al.* used their earlier dataset (# = 1318) for machine learning-assisted polymer design[84] and used failure data to improve PCE prediction.[85] Liu *et al.*[86] created a dataset of 157 non-fullerene ternary OSCs to predict PCE by using FMO of donor, acceptor, and the third component as input descriptors. Ma *et al.*[51] created datasets of 351 polymer : NFA OSCs for optimization of donor–acceptor combination. Sharma *et al.*[87] manually collected a dataset of 154 polymer : NFA OSCs for prediction of non-radiative voltage losses. To predict the PCE of P3HT donor-based OSCs, Wang *et al.*[88] created a dataset of 283 NFAs from the literature with single donor P3HT. Similarly, Wang *et al.*[89] conducted a study on 265 NFAs with single donor PTB7-Th. Wang *et al.*[90] collected a dataset of 164 NFSMAs from the literature with PBDB-T as the donor. In 2022, Lu *et al.*[91] created a dataset of 717 NFA based organic solar cells for PCE prediction.

Aspuru-Guzik *et al.*[92] published a collection of 350 organic small molecules and polymers from the literature that were used as p-type materials in OSCs. This model has been widely used to train QSPR models.[54,93–95] To build structure–property relationship, Sun *et al.*[96] in 2019 developed a dataset of 1719 OSCs with a mix of FA and NFA. For multicomponent materials optimization in BHJ OSCs, Troisi *et al.*[97] in 2019 gathered a dataset of 320 D : A combinations for PCE prediction. In 2019, Kettle *et al.*[98] created a dataset of ~1900 OSCs with corresponding device structure, performance, and stability. Kettle *et al.* later used their dataset to study enhancing stability[99] and understanding the trade-offs between device performance, stability, and environmental impact.[100] To study the effect of increasing descriptor set size, Troisi *et al.*[46] in 2020 created a dataset of OSCs with a mix of FA and NFA. Later in 2022, Troisi *et al.*[101] used the same dataset (# = 566) for PCE prediction of completely new molecule families (extrapolation) by exploring different cross validation techniques. In 2020, Lee[102] gathered the first dataset for 70 tandem OSCs (both conventional and inverted) for PCE prediction. In 2021, Sun *et al.*[103] gathered a dataset of 1758 donor materials (mix of polymers and small molecules) and tested their novel fingerprinting technique for expressing 6180 different fragments (bits). In 2022, Hutchinson *et al.*[104] created a dataset of 84 OSCs and used simplified time-dependent density functional theory (sTD-DFT) for speeding up the calculations by 2–3 times. A summary of datasets used for ML studies in OSCs is given in Table 1.

2.2. Data preparation

After gathering a dataset, suitable set of descriptors are required for prediction of a specific target property. Descriptors used for machine learning applications should be easily accessible so that predictions for unexplored materials or combinations could be quick after saving the model. The solution is to use descriptors that are directly calculated from the chemical structure. In OSC studies, the commonly used input descriptors are microscopic properties (MP), molecular descriptors (MD), molecular fingerprints (FP), frontier molecular orbitals (FMO), and images.

From photon absorption to transport of charge carriers to their respective electrodes, there are a number of microscopic properties that highly influence the PCE of OSCs.[50] Some of the microscopic properties that have been used in ML studies for OSCs are charge carrier mobility, optical bandgap, electron–hole binding energy, and many more. However, they are expensive to

Table 1 Summary of datasets used for ML studies in OSCs

Source	Donor	Acceptor	Data size	Descriptors	Method	Year published	Ref.
Literature	SM	FA	280	MP	Regression	2018	50 and 65
Literature	Polymer	FA	~1200	MP, FP	Regression, classification	2018	66 and 68
Literature	SM	FA	249	MP, FP	Regression	2019	69
Literature	SM	FA	300	MP	Regression	2019	70 and 71
Literature	SM, polymer	FA	124	FMO	Regression	2019	72 and 73
HCEP	Polymer	NFA	51 000	Scharber model	Regression, classification	2017	55, 61, 105 and 106
Literature	Polymer	NFA	~500	FP	Regression	2019	67
Literature	Polymer	NFA	565	FP	Regression	2020	79
Literature	SM, polymer	NFA	135	MP, FMO	Regression	2020	80, 81
Literature	Polymer	NFA	566	MP, FP	Regression	2021	82
Literature	Polymer	NFA	1318	MP, FP, MD	Regression	2021	83–85
Literature	Polymer	NFA	157	FMO	Regression	2021	86
Literature	SM, polymer	NFA	154	MP, FP, MD	Regression	2021	87
Literature	P3HT	NFA	283	MD	Regression, classification	2021	88
Literature	PTB7-Th	NFA	265	MD, DFT	Regression	2022	89
Literature	PBDBT	NFA	164	MD	Regression	2021	90
Literature	Polymer	NFA	351	MP, FP, DFT	Regression	2021	51
Literature	SM, polymer	NFA	717	MD	Regression	2022	91
HOPV	SM, polymer	FA, NFA	350	MP, DFT	Regression	2016	54 and 92–95
Literature	SM, polymer	FA, NFA	1719	FP, MD, Image	Classification	2019	96
Literature	SM, polymer	FA, NFA	320	MP, FP	Regression	2019	97
Literature	SM, polymer	FA, NFA	~1900	Device structure	Regression	2019	98–100
Literature	SM, polymer	FA, NFA	566	MP, FP	Regression	2020	46 and 101
Literature	—	FA, NFA	70	FMO	Regression	2020	102
Literature	SM, polymer	—	1758	FP	Regression	2021	103
Literature	Polymer	FA, NFA	84	MP, DFT	Regression	2022	104

calculate but provide more realistic and accurate measures for organic materials.

To convert the chemical structure of OSC materials into a machine-readable format, simplified molecular input line entry system (SMILES) strings[107] are used. Using SMILES strings, molecular descriptors and molecular fingerprints are generated by using various open-source packages. A comprehensive set of chemical descriptors must be used instead of SMILES strings to describe molecules in order to create a powerful machine learning model.

A molecular descriptor is a machine-readable representation of the information contained within a molecule. Atomic count, atomic type, and molecular weight are all examples of zero-dimensional (0D) descriptors because they make no inferences about topology or atom connectivity. Chemical fragment kinds and their count can be described using one-dimensional (1D) descriptors. Alternatively, two-dimensional (2D) descriptors characterize molecular topology and chemistry. Three-dimensional (3D) descriptors, which take into account conformational information like molecular volume and partial surface charges, capture geometrical information as well.[108]

A molecular fingerprint is the representation of some known structural property of a molecule. When the expected data structure is available, the corresponding bit is set to 1 (ON); otherwise, it is left at 0 (OFF). More bits mean more structural information and could be further used for comparing the structural similarity of two molecules.

FMOs (D_{HOMO}, D_{LUMO}, A_{HOMO}, A_{LUMO}) of donors and acceptors have a direct impact on the charge carrier dynamics in OSCs. With these values, $LUMO_{offset}$, $HOMO_{offset}$, and $Donor_{HOMO}$–$Acceptor_{LUMO}$ are also taken into account to predict better results. For prediction of ternary OSC photovoltaic properties, FMOs of donor, acceptor, and the third component are simultaneously used as input descriptors.[86] Similarly, for tandem OSCs, FMOs for both cells (bottom sub-cell near ITO and top sub-cell near the metal electrode) are used as input descriptors for ML models.[102]

The material structure image or the structure encoded image could be directly used by convolutional neural network (CNN) models for materials property prediction or device property prediction. Models can achieve high accuracy by using images as input, but much larger datasets are required.

Since not all material qualities are relevant for all performance indicators, the list of descriptors that will uniquely define the data will depend on the purpose or information to be retrieved. Before feeding input data into ML models for training, raw data need to be transformed into data that can be effectively used for supervised learning and this process is called feature engineering. The steps involved in the feature engineering process are (1) imputation (handling missing values), (2) encoding categorical variables, (3) handling outliers, (4) scaling (min–max scaling and standardization scaling), and (5) creating new descriptors (from domain knowledge). With the increase in the number of descriptors, the number of data points should be also increased for effectively training a ML model. Since in OSCs, large experimental datasets are not available, high dimensional datasets may confuse ML models, and dimensionality reduction becomes crucial. Input descriptors that are highly correlated are not relevant for the model and are usually dropped based on high correlation coefficient and variance threshold. The recursive feature elimination (RFE) technique is preferably used for this purpose. Principal component analysis (PCA)[109] is a well-known instance of feature extraction, a technique used to reduce dimensionality by constructing a smaller collection of new descriptors from the original list. After applying such feature engineering techniques, the dataset becomes ready to train ML models.

2.3. Model training

ML implementations are classified into three categories: supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning, the model learns the functional mapping between input descriptors and output values. Both regression and classification problems come under supervised learning and is by far the most common type of ML in the field of OSCs. Unsupervised learning models are used to draw inferences from datasets that consist of input data but do not have labeled responses. Lastly, reinforcement learning is a method that simulates the way in which people learn by interacting with their surroundings. This method allows an algorithm to improve its performance on a variety of tasks by receiving feedback in the form of rewards or punishments.

The goal of classification is to predict class labels from the input data. In binary classification, there are only two class labels; if class labels are more than two, it is called multiclass classification. Classification of organic compounds with a high PCE in OSCs has been used in many studies. Because classification methods provide discrete results, a metric is required to compare the distinct classes. Classification metrics measure a model's performance and tell you how excellent or poor the classification is, but each evaluates it differently.

In the regression task, we try to predict a continuous dependent variable using a set of independent variables. For example, photovoltaic properties, such as PCE values, are mostly predicted. Prediction error is used to quantify the success of a model in regression situations. Prediction error, also known as residuals, is the discrepancy between the observed and expected values. The goal of the regression model is to find the best line fit that minimizes the discrepancy between the predicted and observed values. Researchers are actively working to develop a more precise model based on a variety of usability considerations. To gauge the success of a new regression model, it is customary to benchmark it against previously developed models using a set of accepted metrics. Most commonly, these ML architectures for OSC studies are directly accessed from the scikit-learn python package.[110]

Quantitative structure–property relationship (QSPR) models are being used to understand the hidden relationship between materials and their photovoltaic properties.[52] A relationship between the structure of materials and their target property could be obtained by generating feature importance[53] and visualization of decision tree.[71] Various ML methodologies have been used specifically on OSCs for QSPR.[111] Model training is done by selecting a suitable cross-validation method for getting an optimal set of hyperparameters for the selected ML model. Generally, K fold cross-validation is used, but in cases like OSCs where available experimental datasets are not large, leave-one-out-cross validation (LOOCV) is often used. There are currently a plethora of resources available to help with the creation of ML models for use in the materials science field. There has been an explosion of specialized open-source ML software libraries for materials research such as scikit-learn,[112] TensorFlow,[113] and PyTorch.[114]

2.4. Evaluation metrics

A model's performance may be monitored and measured using metrics (during training and testing). In any machine learning pipeline, performance indicators, also called evaluation metrics, play a significant role. Evaluation metrics are used to quantify performance and compare ML models for classification or regression. We have restricted our focus to supervised learning evaluation metrics.

The accuracy (A) metric simply measures how often the classifier predicted correctly. The Confusion Matrix is a table-based representation of the ground-truth labels *versus* the model's predictions. For each row in the confusion matrix, instances in a predicted class are represented, whereas occurrences in a real class are represented in each column. True Positive (TP) denotes the number of positive class samples correctly classified by the model. True Negative (TN) signifies the number of negative class samples classified correctly by the model. False Positive (FP) denotes the number of negative class samples predicted incorrectly by the model. False Negative (FN) signifies the number of positive class samples predicted incorrectly by the model.

The Confusion Matrix is not technically a performance statistic, but it serves as a foundation for evaluating other metrics such as precision (P) and recall (R). When False Positives are of more concern than False Negatives, precision is helpful. Precision is the ratio of True Positives to total number of predicted positives. Recall or Sensitivity is a useful metric in cases where False Negative is of higher concern than False Positive. Recall is the ratio of true positives to total number of actual positives. The F1 score is the harmonic mean of precision and recall. The F1 score is primarily used for comparing different classifier models. The false positive rate (FPR) is defined as the ratio of False Positives to the total number of actual negatives. Specificity is the ratio of true negatives to total number of actual negatives. All major classification evaluation metrics are given in Table 2.

The receiver operating characteristic (ROC) curve is a probability curve that shows how well a classification model works at all classification thresholds by plotting two parameters: the true positive rate (TPR) and the false positive rate (FPR). The area under the ROC curve (AUC) is a measure of how well a model can tell the difference between two classes. The better the model can guess 0's are 0's and 1's are 1's, the higher the AUC.

Some of the most common measures of evaluation for regression models are outlined here. The Pearson correlation coefficient (r) is the most widespread measure for measuring linear correlation. It measures the strength and direction of the relationship between two variables and ranges between -1.0 and $+1.0$. The R^2 score is defined as the proportion of variance in a dependent variable that is predictable from the independent variables. It is also known as ''coefficient of determination''. Mean absolute error (MAE) is one of the most used evaluation metrics and is simply calculated as absolute difference between actual and predicted values. Mean absolute percentage error (MAPE) is the percentage equivalent of MAE normalized by true observations. Mean squared error (MSE) is similar to MAE, but the error is squared here. Root mean squared error (RMSE) is the most widely used evaluation

Table 2 Metrics and their equation used for evaluating the performance of classification and regression ML models

Classification metric	Equation	Regression metric	Equation
Accuracy (A)	$\dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	r	$\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$
Precision (P)	$\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	R^2	$1-\dfrac{\sum_{i=1}^{n}(x_i-y_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Recall (R)/sensitivity/true positive rate (TPR)	$\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	MAE	$\dfrac{1}{n}\sum_{i=1}^{n}\lvert x_i-y_i\rvert$
F1	$2\dfrac{\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}}$	MAPE	$\dfrac{1}{n}\sum_{i=1}^{n}\left\lvert\dfrac{x_i-y_i}{x_i}\right\rvert\times 100$
FPR	$\dfrac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$	MSE	$\dfrac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}$
Specificity/true negative rate (TNR)	$\dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$	RMSE	$\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}}$

metric. It is simply the square root of MSE. Equations for these evaluation metrics are provided in Table 2. In Table 2, x_i denotes the actual value, $\bar{x}$ denotes the mean of actual values, y_i denotes the predicted value, and $\bar{y}$ denotes the mean of predicted values.

2.5. Discovery

Structural flexibility of organic semiconductors provides vast chemical search space for OSC materials, and ML is being utilized for this purpose with full potential. Trained ML models are used for high-throughput screening of novel OSC materials that the models have never seen during the training phase. High-performing materials predicted by the ML models are screened out, and their predicted property is validated experimentally.

3. Review of ML in OSC research

3.1. FA based OSCs

In 2018, Ma *et al.*[50] gathered a dataset of 280 small molecule OSCs with 270 distinct small molecule donors and two distinct acceptors ($PC_{61}BM$ and $PC_{71}BM$). Five ML models (LR, k-NN, ANN, RF, GB) for PCE prediction were employed with 13 microscopic properties as inputs for ML models. The 13 microscopic properties used as inputs for ML models are (1) number of conjugated atoms in the main conjugation path of the donor molecule ($N^{\mathrm{D}}_{\mathrm{atom}}$), (2) polarizability, (3) energetic difference of LUMO and LUMO−1 (Δ_{L}), (4) energetic difference of HOMO and HOMO−1 (Δ_{H}), (5) vertical ionization potential of donor molecules (IP(ν)), (6) reorganization energy for holes in donor molecules (λ_{h}), (7) hole–electron binding energy in donor molecules (E_{bind}), (8) energetic difference between LUMO of donor and LUMO of acceptor ($E^{\mathrm{DA}}_{\mathrm{LL}}$), (9) energetic difference between HOMO of donor and LUMO of acceptor ($E^{\mathrm{DA}}_{\mathrm{HL}}$), (10) energy of transition to singlet excited state with largest oscillator strength (E_{g}), (11) change in dipole moment in going from ground state to first excited state for donor molecules (Δ_{ge}), (12) energy of electronic transition to lowest-lying triplet state (E_{T1}), and (13) energetic difference between LUMO and LUMO+1 of acceptor ($\Delta^{\mathrm{A}}_{\mathrm{L}}$). The GB model performed the best for the test set (r = 0.79) (Fig. 3(a)) and using LOOCV (r = 0.76) (Fig. 3(b)). As shown in Fig. 3(c and d), descriptor importance revealed that for both ensemble techniques GB and RF, out of the 13 descriptors, hole–electron binding energy (E_{bind})[115] is the most important descriptor.

Saeki *et al.*[66] designed a polymer for OPV using machine learning for the first time. Despite the advances in ML, data-driven approaches for OPV are not performing that well. In this work, the authors have used supervised algorithms (RF and ANN) for screening potential polymer–fullerene OPV. The authors conducted a study on ∼1200 polymer–fullerene data collected from the literature (500 papers). Using the random forest model, they achieved a correlation coefficient (r) of 0.6 to 0.7. Various parameters have been used in this study such as M_{w}, E_{g}, FMO, and fingerprints (MACCS[116] and ECFP6[117]). In the next step the polymer design scheme is used as depicted in Fig. 4. By using the 2.3 million HCEP dataset, 1000 molecules were selected, and MACCS fingerprints were used to train the classification model for PCE. The model gave an accuracy of 48%. Based on synthetic feasibility, they selected one polymer with predicted PCE (5% to 5.8%) but got only 0.53% PCE experimentally. The reasons for poor results by the authors are (1) less dataset and (2) poor performance of the Scharber model. Non-availability of exact

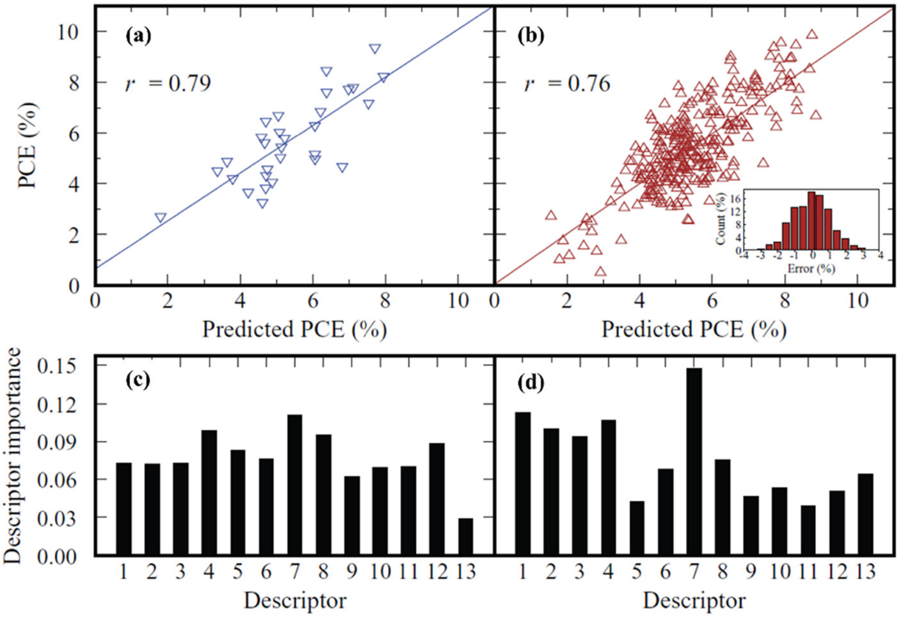

Fig. 3 (a) Theoretically predicted *versus* experimental PCE for the testing set (30 molecules). (b) All data points using the LOO cross validation technique for the GB model. (c) The descriptor importance for the GB. (d) The descriptor importance for the RF. Descriptors are in the following order: (1) $N^{\mathrm{D}}_{\mathrm{atom}}$, (2) polarizability, (3) ΔL, (4) ΔH, (5) IP(ν), (6) λh, (7) Ebind, (8) $E^{\mathrm{DA}}_{\mathrm{LL}}$, (9) $E^{\mathrm{DA}}_{\mathrm{HL}}$, (10) energy of transition to singlet excited state with largest oscillator strength (E_{g}), (11) change in dipole moment in going from Δ_{ge}, (12) ET1, and (13) $\Delta^{\mathrm{A}}_{\mathrm{L}}$. Reproduced with permission[50] Copyright, 2018, John Wiley and Sons.

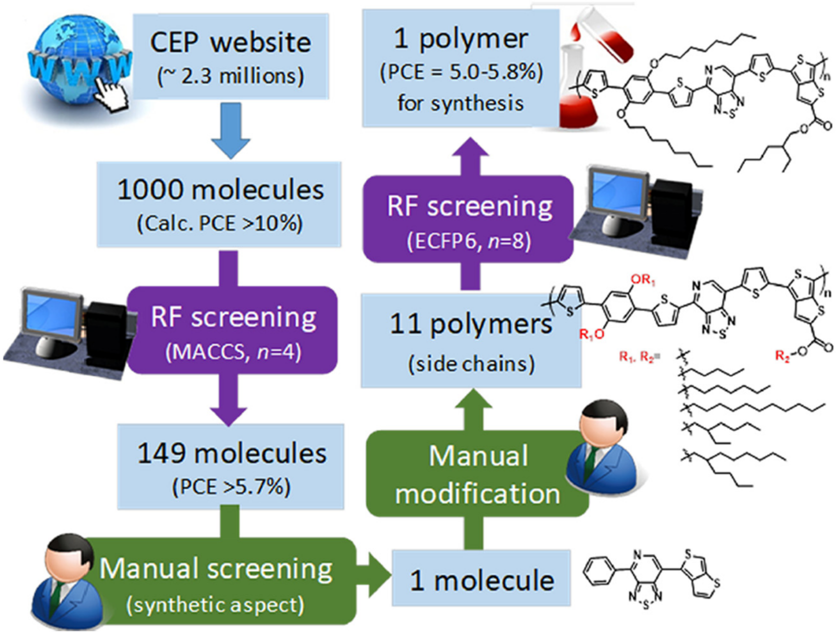

Fig. 4 Scheme of polymer design by combining the RF screening and manual screening/modification. The picked-up molecule or polymer in each stage is shown. Reproduced with permission[66] Copyright © 2018 American Chemical Society.

processing conditions, including film thickness, solvent, solvent additive, p/n blend ratio, and thermal solvent annealing, is a significant issue for this study. The authors also focused on the unavailability of essential data such as (1) purity of the polymer, (2) surface free energy, (3) polymer orientation, and (4) BHJ morphology associated with miscibility.

Chen[68] used a dataset of 1203 polymer–fullerene OSCs by Saeki *et al.*[66] for virtual screening using RF and SVM models. Using RDKit,[118] morgan fingerprints[117] were calculated as input descriptors, and the use of morgan radius further improved the model accuracy. Predictions of these two models were further ensembled to achieve more accurate results (r = 0.653).

Using the concepts of design of experiments (DoE),[119] one may test and optimize several variables at the same time. Minimizing experimental effort while maximizing experimental information output is meant by DoE, a strategy for conducting experiments. In 2018, Buriak *et al.*[47] demonstrated the ML model and DOE integration to create meaningful multidimensional maps for insights on PCE. For DoE on the PCDTBT : PCBM solar cell, four initial parameters with four levels each were chosen: (1) donor weight percentage, (2) total solution concentration, (3) bulk heterojunction spin cast speed, and (4) processing additives. A complete analysis would require 4^4 = 256 experiments; by using fractional factorial design only 16 experiments were performed. With variance analysis of all the parameters, the contribution of the "processing additive" was low and hence was dropped for further analysis. These data were fit using support vector machine (SVM) with radial basis function (RBF) kernel to determine PCE. Three-dimensional maps were generated to visualize PCE approximation in three-parameter space.

Troisi *et al.*[69] in 2018 used the Scharber model on a dataset of 249 organic donor–acceptor pairs, where acceptors are limited to only C_{60}, $PC_{61}BM$, or $PC_{71}BM$. High accuracy was attained using the Tanimoto similarity index between donors and the Euclidean distance of electronic properties for PCE predictions. Prediction results were obtained utilizing both topological and electronic descriptors with KRR (r = 0.68) and k-NN (r = 0.61) algorithms. The authors concluded that the Scharber model performs poorly when studied on a collected dataset with r = 0.38 for V_{oc}, 0.17 for J_{sc}, and 0.18 for PCE. Ma *et al.*[70] (2019) created a dataset of 300 OSCs (small molecule donor and fullerene acceptor). More than 10K molecules were generated for creating a search space by using 32 unique molecular building blocks (donor, acceptor, Π-spacer, and end-capping groups) and 17K DFT calculations. Five machine learning models were trained on 250 data points using 10-fold cross-validation, and the results were compared. The best results were achieved by GBRT (r = 0.80) and the worst results by the Scharber model (r = 0.14). The authors gave the following reasons for the failure of the Scharber model: (1) oversimplified model and (2) the limitations of DFT in estimating FMO. The authors identified important moieties using the GBRT model. They filtered out 126 potential molecules with predicted PCE > 8% and suggested synthesizing materials and fabricating the devices.

Ma *et al.*[71] used an earlier dataset of 300 fullerene-based OSCs[70] and used two ML models (RF and GBRT) to predict J_{sc}, V_{oc}, FF, and PCE. Special attention is given to J_{sc} and V_{oc} prediction because of potential commercial applications: water splitting for high solar to fuel energy conversion (high V_{oc} is required) and solar-window application (high J_{sc} is required). 13 molecular properties were used as input descriptors for ML models and also inspected visually by using decision tree for all target variables (J_{sc}, V_{oc}, FF, and PCE). The GBRT model gave the best results for all four cases J_{sc} (r = 0.66), V_{oc} (r = 0.67), FF (r = 0.71), and PCE (r = 0.78).

Ternary OSCs are made by introducing a third component into the binary blend. Adding a third component in the blend increases light harvesting, reduces energy loss, and creates more D/A interfaces, facilitating exciton dissociation more effectively. In 2019, Lee[72] created a dataset of 124 fullerene derivative-based ternary OSCs (published during 2012–2019). The dataset was manually collected from the literature (> 100 papers) and consisted of FMOs of donor, acceptor, and the third component, and their corresponding PCE. Using this dataset, five ML models (RF, GB, k-NN, LR, and SVR) were compared to predict PCE, and the best results were achieved with RF (R^2 = 0.77) followed by GB (R^2 = 0.73). The authors also studied a two-class classification model on the same dataset, class1 (PCE > 9%) and class2 (PCE < 9%). Again, the RF model performed the best with a test accuracy of 0.76, and with further hyperparameter optimization, the test accuracy improved to 0.855. Lee[73] used an earlier dataset[72] (# = 124) for the prediction of open-circuit voltage (V_{oc}) in ternary OSCs using ML models (RF and SVR). FMOs of donors, acceptors, and the third component were taken as input descriptors. The RF model performed the best with R^2 = 0.77 for the test set and was further used to generate feature importance scores. According to the model, HOMO and LUMO of donor are the two most important factors that influence V_{oc}. Using a contour plot, the author visualized optimal energy level alignment rules for donor and the third component.

Small datasets are not ideal for training a neural network; in the case of OSCs, large experimental datasets are not available. MacKenzie *et al.*[120] generated a dataset of 20 000 devices using the Shockley–Read–Hall based drift-diffusion model using gpvdm. For all the devices, dark and light current–voltage curves were simulated using randomly assigned electrical parameters such as trap densities, recombination time constants, carrier trapping states, energetic disorder, and parasitic resistance. They demonstrated their method for getting optimal surfactant choice and annealing temperature in terms of charge carrier dynamics in P3HT, PBTZT-stat-BDTT-8, and PTB7 based OSCs with PCBM acceptors. All these electrical parameters were included in the output layer node of the neural network as depicted in Fig. 5. After model training the neural network is used to predict electrical parameters using dark and light current–voltage curves as input.

Goharimanesh *et al.*[65] in 2022 used 240 small molecule OSC data from the dataset of Ma *et al.*[50] The performance of six ML models (LR, k-NN, ANN, RF, GB, and XGB) was compared for the prediction of PCE. Expensive quantum chemical descriptors were used as input (calculated with the Gaussian 09 package), and ANN performed the best (r = 0.79) on the training set. The authors also designed a new technique for mapping structure–property relationships by combining the Taguchi design of experiments (TDOE) approach and ML. The complete workflow of this study using the TDOE approach is represented in Fig. 6. A PCE map represented model interpretability for the following descriptors studied: energetic difference between LUMO+1 and LUMO of donor (Δ_L), energetic difference between HOMO and HOMO−1 of donor (Δ_H), optical gap (Δ_{HL}), hole reorganization energy (λ_h), exciton binding energy (E_b), and LUMO band offset (Δ_{LUMO}).

3.2. NFA based OSCs

Aspuru Guzik *et al.*[61] in 2017 created a dataset of 51 000 NFAs based on tetraazabenzodifluoranthenes (BFIs), diketopyrrolopyrroles (DPPs), perylene diimides (PDIs), benzothiadiazole (BT),

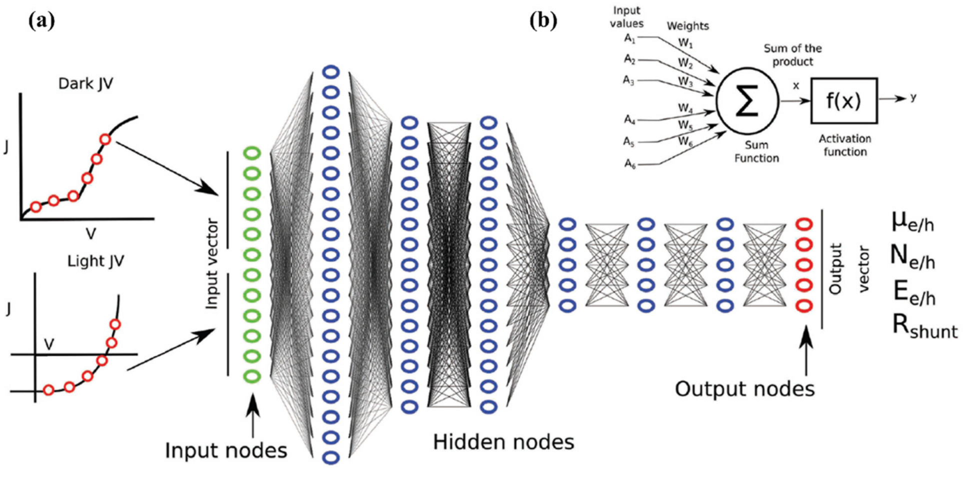

Fig. 5 A diagram of the neural network used to extract material parameters from the data within this paper; the actual network used had ten times more neurons in each hidden layer than the diagram depicts but otherwise the same structure. Visible on the left hand side of the image are the experimental (or simulated) data, with the red dots on the curves representing the points at which the curves were sampled to form input vectors for the neural network. (b) In the diagram a light and a dark JV curve are each being sampled at 6 places to provide 12 data points to the neural networks 12 input nodes. Any number or combination of experimental measurements can be placed on the input to the network; one simply has to extend the number of input neurons, and retrain the network. The neural network itself has red input nodes, blue hidden layers, and green output nodes. Each output node corresponds to a device/material parameter such as charge carrier mobility or trap density. Inset: A single neuron. Reproduced with permission[120] Copyright © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

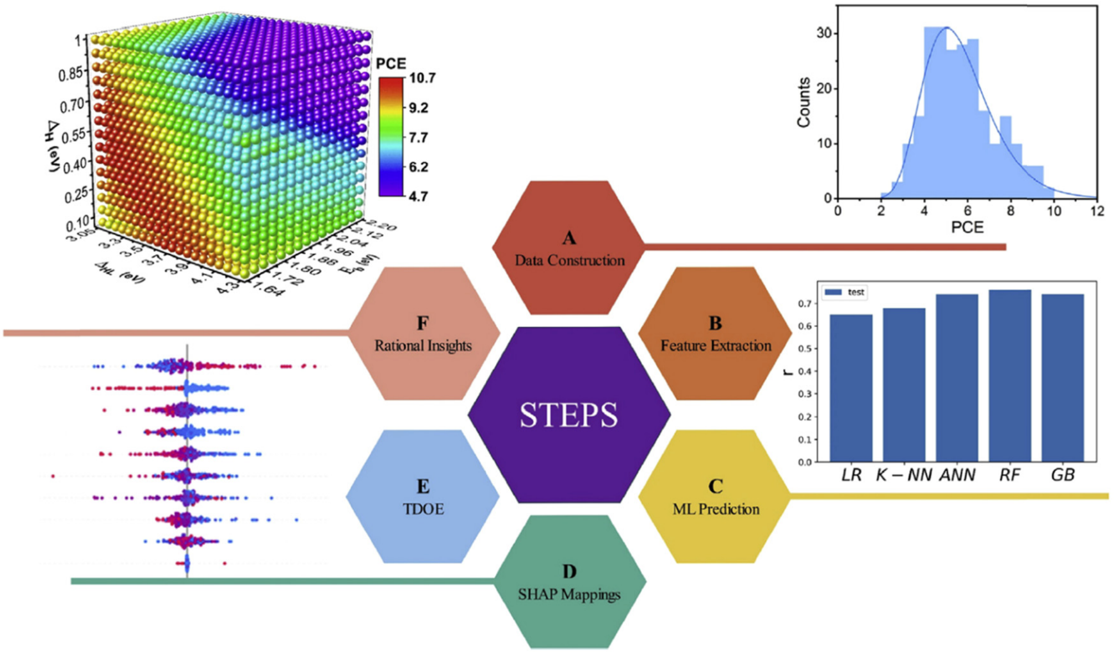

Fig. 6 Steps of the research in this paper: (A) data construction, (B) feature extraction, (C) ML prediction, (D) SHAP mappings, (E) optimization by TDOE, and (F) rational insights. Reproduced with permission[53] Copyright © 2022 Elsevier Ltd.

and fluoranthene-fused imides. Since the Scharber model was used to predict the PCE of NFAs and poly[*N*-90-heptadecanyl-2,7-carbazole-*alt*-5,5-(40,70-di-2-thienyl-20,10,30-benzothiadiazole)] (PCDTBT) as the donor material, poor results were obtained while comparing true and predicted PCE (r = 0.43 and R^2 = 0.11).

To discover novel active materials for OSCs, generative models[121] are being used effectively. Generative models generate new molecules by creating a latent space from molecular information and then manipulating the latent space. For OSCs, novel polymer donors[54,63] and NFAs[55] are generated using generative models. Hada *et al.*[63] developed a model for high throughput screening of thiophene-based D–A polymers. They automated the process in three steps: (1) polymer generation, (2) orbital energy estimation by Hückel-based model, and (3) photovoltaic property calculation. Using donor and acceptor units, 380 polymers were generated and studied, followed by identification of promising acceptor units for photovoltaic materials. The results are not significant in terms of PCE because the Scharber model solely uses orbital energies.

In recent years, CNNs have established themselves as the standard framework for computer vision tasks. Then the training model is utilized for regression or classification task. Because of its capacity to build and evaluate features in image-like inputs, CNN has become extremely popular. Zhao *et al.*[55] in 2019 used CNN for the generation of new acceptors and their structure–property prediction. The authors have demonstrated the generation of new molecules by CNN and controlling the diversity of the molecules by variation in the number of convolution layers. Post molecule generation, various descriptors are used for property prediction such as extended connectivity fingerprints, molecular graphs, bag of bonds, coulomb matrix, and SMILES strings. The attention mechanism is adopted to get character-specific weights in SMILES strings. The model is used for prediction of HOMO, LUMO, and PCE. The authors have used a dataset from Aspuru Guzik *et al.*'s (# = 51 000) dataset.[61] For this study, all the molecules with PCE $<$ 0.5% were removed, and 24 000 molecules were chosen randomly. Out of this, 20 000 were used for training, 2000 for evaluation, and 2000 for testing. The donor used for this study is fixed, and the Scharber model was used for calculating PCE.

In 2019, Sun *et al.*[105] used CNN on the HCEP dataset for the prediction of PCE. Without any prior transformation, the CNN model uses chemical structure images of donor materials as input. The dataset is classified into two categories of PCE (0–4.9% and 5–9.9%). The classification model achieved an accuracy of over 90%. The impact of data size on model accuracy is also studied. The major drawbacks of this study are (1) donor molecules being used today are much more structurally complex and bigger in size compared to those in the HCEP dataset and (2) PCE is calculated using the Scharber model, and the Scharber model estimates PCE using DFT-calculated energy levels.

Using a similar approach as in ref. 66, Wei *et al.*[67] performed a similar study with a 500 polymer : SMA dataset using ECFP6 fingerprints. The authors studied two new polymers (BO2FC8 and BO2FEH) and using the random forest model they predicted PCE (11.2% and 10.9%). They also validated the results with PCE of 11.0% and 6.4% from experimental findings. Interestingly the former gave consistent results with the ML model, while the latter results did not agree with the ML model.

Min *et al.*[79] used five ML models, linear regression (LR), multiple linear regression (MLR), boosted regression trees (BRT), random forest (RF), and artificial neural network (ANN), on 565 polymer donor : NFA pairs to predict PCE.

The dataset was manually collected from the literature (274 papers). 477 D : A pairs were used for training and the best results were achieved by BRT (r = 0.71) and RF (r = 0.7), while the other models performed poorly (LR (r = 0.54), MLR (r = 059), and ANN (r = 0.6)). Based on the structural fragment in donor and acceptor in 565 D : A pairs, the authors created a search space dataset of 3 20 76 000 D : A pairs and predicted their PCE using the trained BRT and RF model. The complete workflow is depicted in Fig. 7. From the predictions of both models, it was observed that for RF, 12.27% of D : A pairs gave a PCE of greater than 11%, while for BRT, 14.15% D : A pairs gave a PCE of greater than 11%. Six D : A pairs were selected for experimental validation with donors (PM6 and PBDB-T) and acceptors (Y-ThCN, Y-ThCH3, and Y-PhCl). These acceptors were selected for experimental validation because they are easily synthesizable with a one-step method. The PCE values obtained in experiments were quite similar to those predicted by BRT and RF.

Lee[80] created a dataset of 135 donor : NFA pairs from literature reviews[122,123] consisting of 117 unique NFA and 30 unique donor materials. FMOs and bandgaps were used as input descriptors to predict PCE using RF and GB regression models. For the test set, the RF model gave the best results (R^2 = 0.80), and the acceptor bandgap was given the highest feature importance. Buriak *et al.*[124] in 2020 used the same ML-DOE approach as in ref. 47 for an all-small-molecule organic solar cell with the donor DRCN5T and acceptors ITIC, IT-M, and IT4F. Their work focused on active layer optimization of OSCs, and four initial parameters were used for DoE: (1) solution concentration of donor and acceptor ink, (2) donor fraction, (3) temperature, and (4) annealing time. Maps were derived for power conversion efficiency to visualize the effect of parameters and find the most optimal combination of parameters. It is important to note that fitting functions like RBF are good for interpolation only and should not be applied for extrapolation tasks.

Saeki *et al.*[82] manually collected a dataset of 566 polymer donor : non-fullerene acceptors and used material property (E_g, HOMO, LUMO, M_w) and ECFP6 fingerprints for prediction of PCE. The dataset was collected until 2018 and does not include high-performing NFA of Y6 or ternary solar cells. The best result (r = 0.85 ± 0.02) was obtained using 5-fold CV and r = (0.85 ± 0.02) with LOOCV. From their earlier polymer : FA dataset (# = 1203)[66] and current polymer : NFA dataset (# = 566), donor and acceptor units were extracted, and a total of 2 00 932 D–A polymers were virtually generated as shown in Fig. 8. For all generated D–A polymers, PCE was predicted with acceptors ITIC and IT-4F. Using the RF model on ECFP6 fingerprints, the authors revealed that feature importance of ECFP6 fingerprints was more than polymer molecular properties, and the model gave similar results when polymer molecular features were removed. Interestingly for IT4F, the 20 polymers with the highest predicted PCE are analogs of PBDB-T. From top predictions, they selected second-rank polymer PBDT(SBO)TzH and prepared a device with IT4F. Its predicted efficiency was 10.5%, but it achieved 3.42% PCE in the experiment, which is quite low. This unsatisfactory performance of the polymer is due to the poor solubility, which is also evident from high PDI. They modified PBDT(SBO)TzH and synthesized PBDTTzH, PBDTTzEH, PBDTTzBO, and PBDTTzHD to improve the solubility. PBDTTzEH achieved a PCE of 7.5%. Although the ML workflow did not suggest

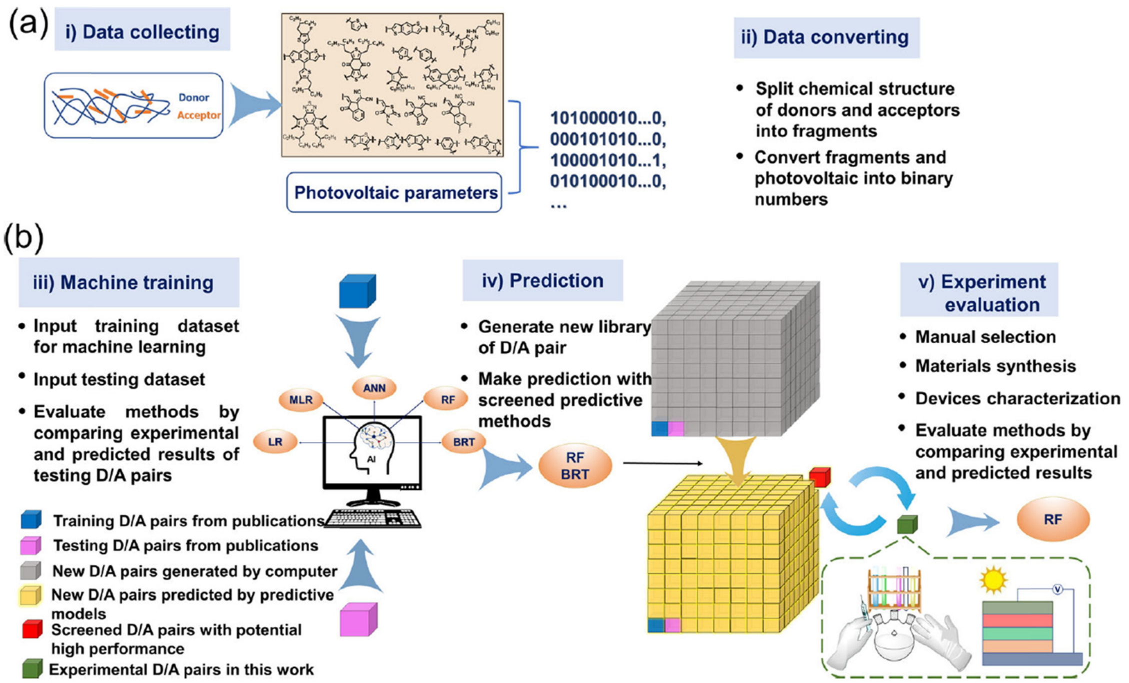
(a)
i) Data collecting
Donor
Acceptor
Photovoltaic parameters
101000010...0,
000101010...0,
100001010...1,
010100010...0,
...
ii) Data converting
• Split chemical structure of donors and acceptors into fragments
• Convert fragments and photovoltaic into binary numbers
(b)
iii) Machine training
• Input training dataset for machine learning
• Input testing dataset
• Evaluate methods by comparing experimental and predicted results of testing D/A pairs
MLR
ANN
RF
LR
BRT
AI
iv) Prediction
• Generate new library of D/A pair
• Make prediction with screened predictive methods
RF
BRT
v) Experiment evaluation
• Manual selection
• Materials synthesis
• Devices characterization
• Evaluate methods by comparing experimental and predicted results
RF
V
🟦 Training D/A pairs from publications
🟪 Testing D/A pairs from publications
⬜ New D/A pairs generated by computer
🟨 New D/A pairs predicted by predictive models
🟥 Screened D/A pairs with potential high performance
🟩 Experimental D/A pairs in this work

Fig. 7 Workflow of building, application, and evaluations of machine learning methods in this work. (a) Scheme of collecting experimental data and converting chemical structures to digitized data. (b) Scheme of machine training, prediction, and method evaluation. Reproduced with permission[79] Copyright © 2020, The Authors, Springer Nature.

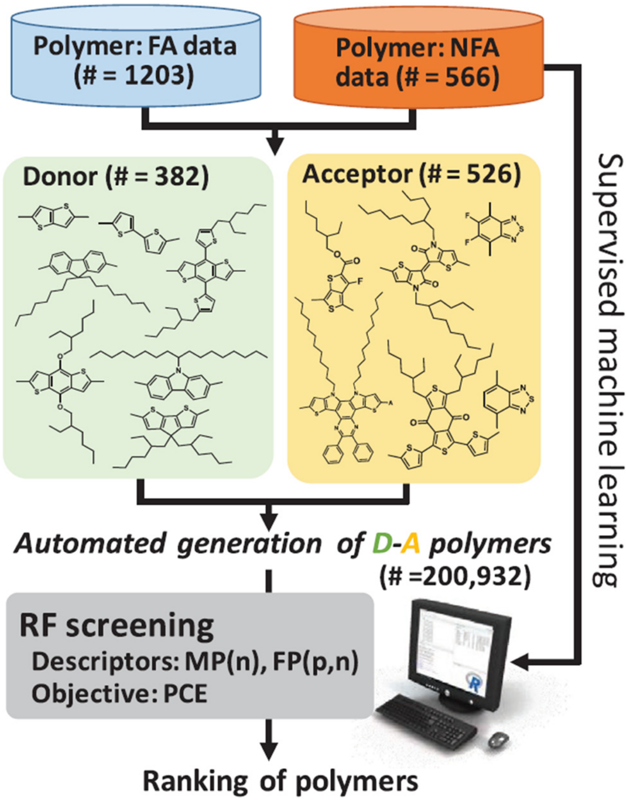
Polymer: FA data (# = 1203)
Polymer: NFA data (# = 566)
Donor (# = 382)
Acceptor (# = 526)
Supervised machine learning
Automated generation of D-A polymers (# =200,932)
RF screening
Descriptors: MP(n), FP(p,n)
Objective: PCE
Ranking of polymers

Fig. 8 Scheme of polymer screening for polymer : NFA OPVs. The number (#) of polymer : FA[66] and polymer : NFA (present work) data points was 1203 and 566, respectively. Donor (D) and acceptor (A) units were extracted from these data after removing overlaps and a total of 200 932 new D–A polymers were virtually generated. Polymer screening by the RF model constructed on the supervised ML of polymer : NFA was applied using MP(n) and FP(n,p) as the descriptors. Reproduced with permission[82] Copyright © 2021 Wiley-VCH GmbH.

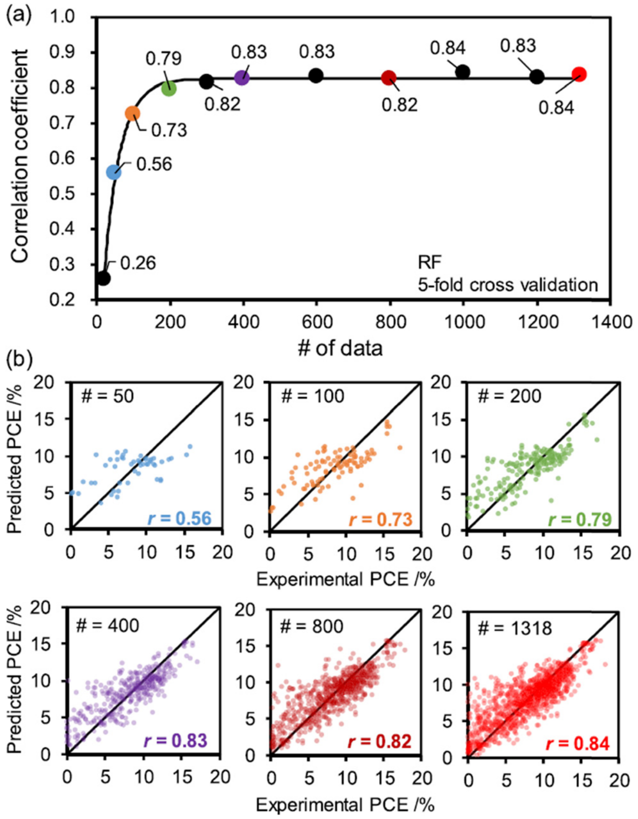

Fig. 9 (a) Effect of the number of data points (#) of NFA-OPV on the correlation coefficient of the RF model (5-fold CV). (b) Plots of predicted and experimental PCE obtained for each #. The color corresponds to the plot in panel (a). Reproduced with permission[83] Copyright © 2021 American Chemical Society.

additives, after using DPE as an additive (0.5 vol%), PCE further improved to 10.1%.

Saeki *et al.*[83] increased their dataset for study from 566[82] to 1318. It is important to note that Y6 and its derivatives are included in the increased dataset. The authors studied the effect of dataset size by using ECFP6 fingerprints, Mordred descriptors, and molecular properties and applying a 5-fold CV. The authors studied the effect of categorizing different NFAs and observed that merging all groups into one dataset improves the overall result. The authors also studied the effect of increasing number of data points on ML predictions, and the results are depicted in Fig. 9. The results reveal that $r = 0.25$ for # = 50 and $r = 0.79$ for # = 200, and saturates at $r = 0.82$ for # = >300. For # = 1318, observed $r = 0.84$.

Saeki *et al.*[84] developed three new polymers (PBDTTzBO, PFSBDTTzBO, and PFBDTTzBO) and predicted their PCE with NFAs (IT-4F and Y6). Using the RF model trained on the 1318 dataset,[83] predicted PCEs of these three polymers with IT-4F (9.93, 11.35, and 11.47%) were found to be in good agreement with experimental validation (5.24, 7.35, and 10.30%). On the other hand, for Y6, an inverse trend was observed between predicted values (9.20, 12.29, and 12.20%) and experimental validation (11.98, 1.57, and 6.53%). The authors mentioned that an inverse relationship in predicted and experimental values of Y6 PCE is due to a small training set of Y6 acceptors (# = 46) in the complete dataset (# = 1318).

Liu *et al.*[86] in 2021 created a dataset of 157 non-fullerene ternary OSCs (published during 2015–2020). Five ML models (RF, XGBoost, KNN, Decision tree, SVM, Ridge regression, MLP) were used to predict ternary OSCs' PCE and compare their results. FMOs of donor, acceptor, and the third component were used as input descriptors. The RF model performed the best for this task. Moreover, fine-tuned classification models were also compared, and the RF model again gave the best results indicating its suitability for such application. The study outcomes make it clear that (1) LUMO of NFA can slightly lower V_{oc} but significantly enhance J_{sc} and (2) V_{oc} could be optimized by shifting the LUMO of the third component slightly up in ternary OSCs.

Compared with other photovoltaic technologies such as perovskite solar cells, OSCs have lower V_{oc} and higher non-radiative voltage loss. Sharma *et al.*[87] created a dataset of 154 unique D : A combinations with their corresponding non-radiative voltage loss. Combinations of molecular descriptors, fingerprints, FMO, and optical bandgap (E_g) were used as inputs for ML models to predict non-radiative voltage loss.

Two molecular descriptors (RDKit[118] and Mordred[125]) and four fingerprints (CDK,[126] MACCS,[116] PubChem,[127] and Morgan[117]) were used in this study. Four ML models (RF, GBR, SVR, and ANN) were used to predict non-radiative voltage loss. The trend of PCE with non-radiative voltage loss was analyzed, and RF obtained the best results by using a combination of FMO, optical bandgap, and RDKit descriptors as input (r = 0.857). The ML workflow for non-radiative voltage loss is depicted in Fig. 10.

Wang *et al.*[88] designed a ML model for P3HT : NFA based OSCs (# = 283). More than 3000 easily synthesizable small molecule acceptors were designed using various building blocks. A dataset of 764 small molecule organic semiconductors was collected from the literature, and their corresponding molecular descriptors were calculated using OCHEM[128] and Chemdes.[129] KNIME[130] and Weka[131] open-source platforms were used to perform ML studies. The authors used a two-step screening approach as demonstrated in Fig. 11. In the first step, FMOs were predicted using the LR model with # = 764 dataset, followed by the screening of molecules whose FMO aligns properly with P3HT. After screening out 500 NFAs, in the second step their PCE is predicted using the SVM regression model. Further for selection of suitable green solvents, the RF model was trained on a dataset of 252 solvents to predict their corresponding Hansen solubility parameters (HSP).

In 2021, Brabec *et al.*[48] created an autonomous materials and device application platform (AMANDA Line One) for high-throughput screening and device fabrication. The study was conducted on PM6/Y6 combination for evaluation of PCE, J_{sc}, V_{oc}, FF, and photostability using Gaussian process regressor (GPR). With this method, the authors reported screening of PM6 : Y6 OSCs with ∼100 processing conditions within 70 hours (including the photostability test) and the highest PCE of 14% with a fully automated fabrication process. The processing

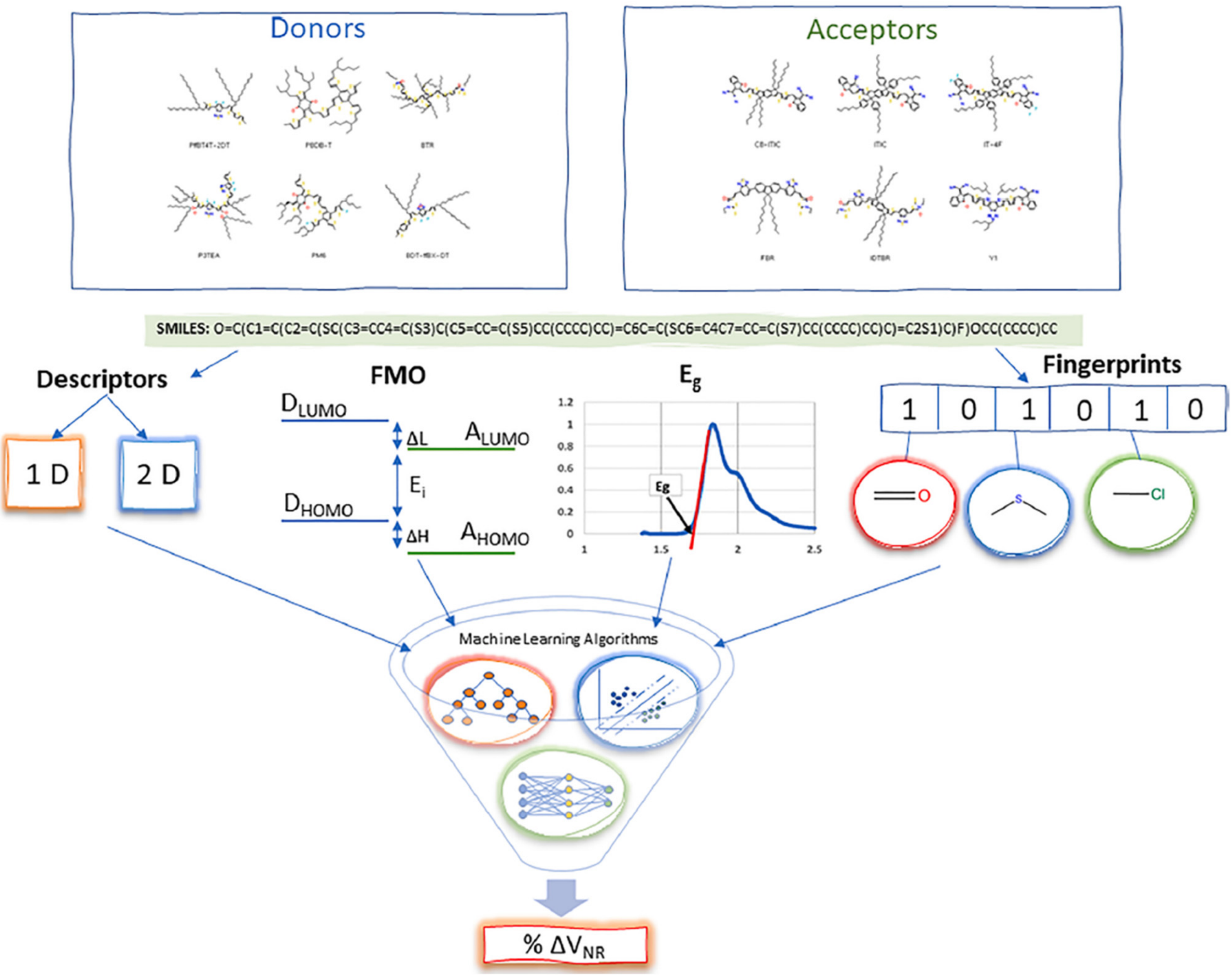

Fig. 10 Machine learning workflow for the prediction of %ΔV_{NR}. Data for 154 unique D : A combinations with reported %ΔV_{NR} are collected from the literature, having 46 distinct donors and 79 distinct acceptors. Reported FMO and E_g values are taken from the literature and then transformed by median values for distinct donors and acceptors. SMILES codes of donor and acceptor molecules are generated by using ChemDraw software. SMILES codes are then used to generate molecular descriptor datasets and molecular fingerprint datasets. Finally, the datasets are scaled and fed into ML models for the prediction of %ΔV_{NR}. Reproduced with permission[87] Copyright © 2021 International Solar Energy Society. Published by Elsevier Ltd. All rights reserved.

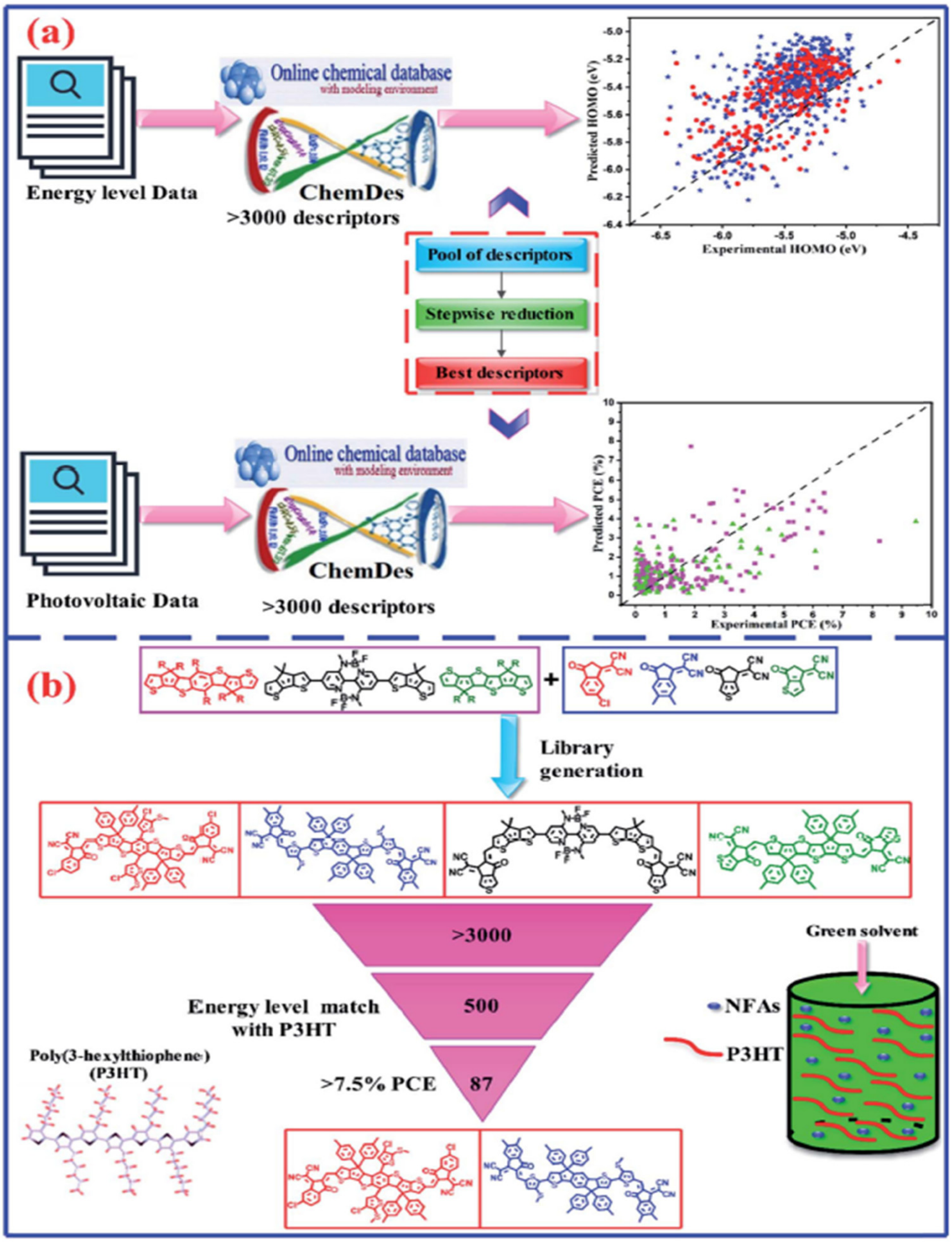

Fig. 11 Brief description of the whole study. (a) Model training. (b) Library generation and screening of potential candidates. Reproduced with permission[88] Copyright, 2021, Royal Society of Chemistry.

parameters include concentration, D : A ratio, active layer annealing temperature and time, spin speed, solvent additive variations in materials and volume, electron transport layer (ETL) material variation, and ETL annealing temperature and time. With this study, photovoltaic parameters and device stability could be predicted using GPR with high accuracy. Also, the PCE and burn-in losses could be improved by using devices with a medium thermal annealing temperature and a thin active layer.

Wang *et al.*[89] collected a dataset of 265 NFAs with the PTB7-Th donor. ML models (RF, k-NN, LR, and SVM) were used for the prediction of FMO, PCE, and absorption maxima. Input descriptors were calculated using Gaussian 09, OCHEM,[128] and Chemdes.[129] The RF model gave the best results for predicting PCE on the test set (r = 0.93). Easily synthesizable building blocks were used to construct more than 5000 novel small molecule acceptors (SMAs). Over 1700 small molecule acceptors were discarded since they didn't have matching energy levels in common with the PBT7-Th. Blue-shifted SMAs were not evaluated any further. Based on the predicted UV/visible absorption maxima, the total number of SMAs was whittled down to 2350. The SMAs were then further scrutinized based on the predicted PCE. Molecular dynamics (MD) simulations were used to investigate more than 100 SMAs with more than 13% PCE as shown in Fig. 12. The Flory–Huggins parameter was used to investigate the mixing behavior of PBT7-Th : SMA blends. A total of 15 SMAs were chosen because of their ability to combine well with PBT7-Th.

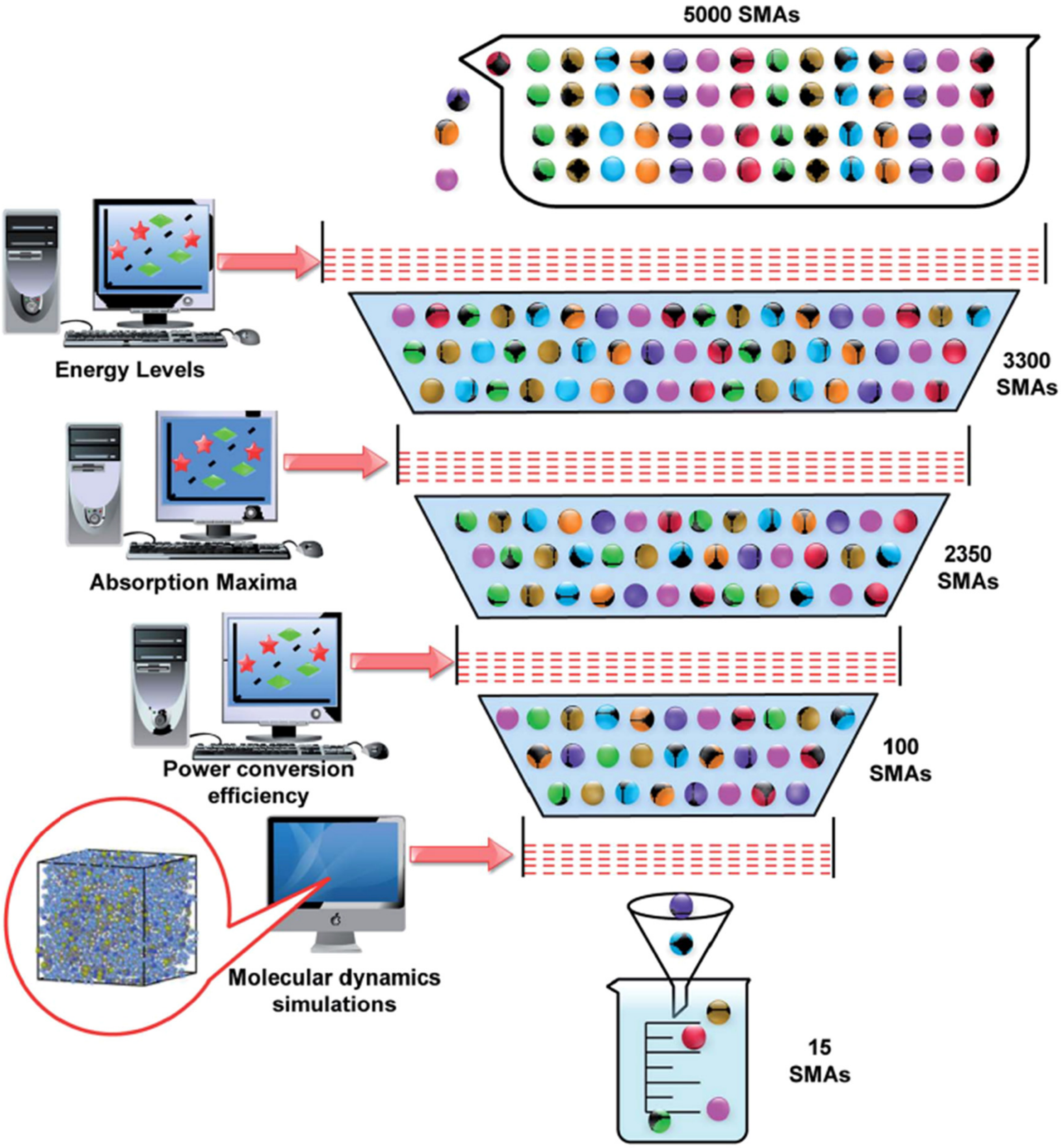

Fig. 12 The screening pipeline to screen the designed SMAs for PBT7-Th : SMA-based OSCs. Reproduced with permission[89] Copyright, 2022, Royal Society of Chemistry.

Ma *et al.*[51] created a reliable QSPR model that integrates chemical descriptors and device requirements and achieves a promising inverse optimization for unexplored potential D : A combinations. The dataset, consisting of 351 D : A combinations (44 polymer donors and 195 NFAs), was collected from the literature. Device specifications, 19 electronic properties (using DFT), and six fingerprints were used as inputs for ML models. For developing the QSPR model, different combinations of these input descriptors were studied. With various combinations of these descriptors, sixty ML models were created with ridge regression, GB, SVR, ANN, and voting ensemble approach (combines predictions of all regression models). Since CDK fingerprints performed the best out of all, it was used for further analysis. The voting model obtained the best results for PCE prediction ($r > 0.8$). This QSPR model was further integrated with high-throughput screening by creating a search space of 1942785 D/A pairs, and D/A pairs with PCE greater than 14% were screened out. The complete workflow is demonstrated in Fig. 13.

Vak *et al.*[49] in 2021 developed a novel research method that enables quick screening of vast parameter space using methods relevant to industries (roll-to-roll processing). The PM6 : Y6 : IT4F ternary system was chosen for the experiment, and 2218 OSCs were fabricated by varying the ratios. This study used two new terminologies: DD (deposition density) and TDD (total deposition density), which refer to the quantity deposited per unit area (in g cm^{-2}) of each substance and their total (the sum of PM6, Y6, and IT-4F) quantities. As a result, they give

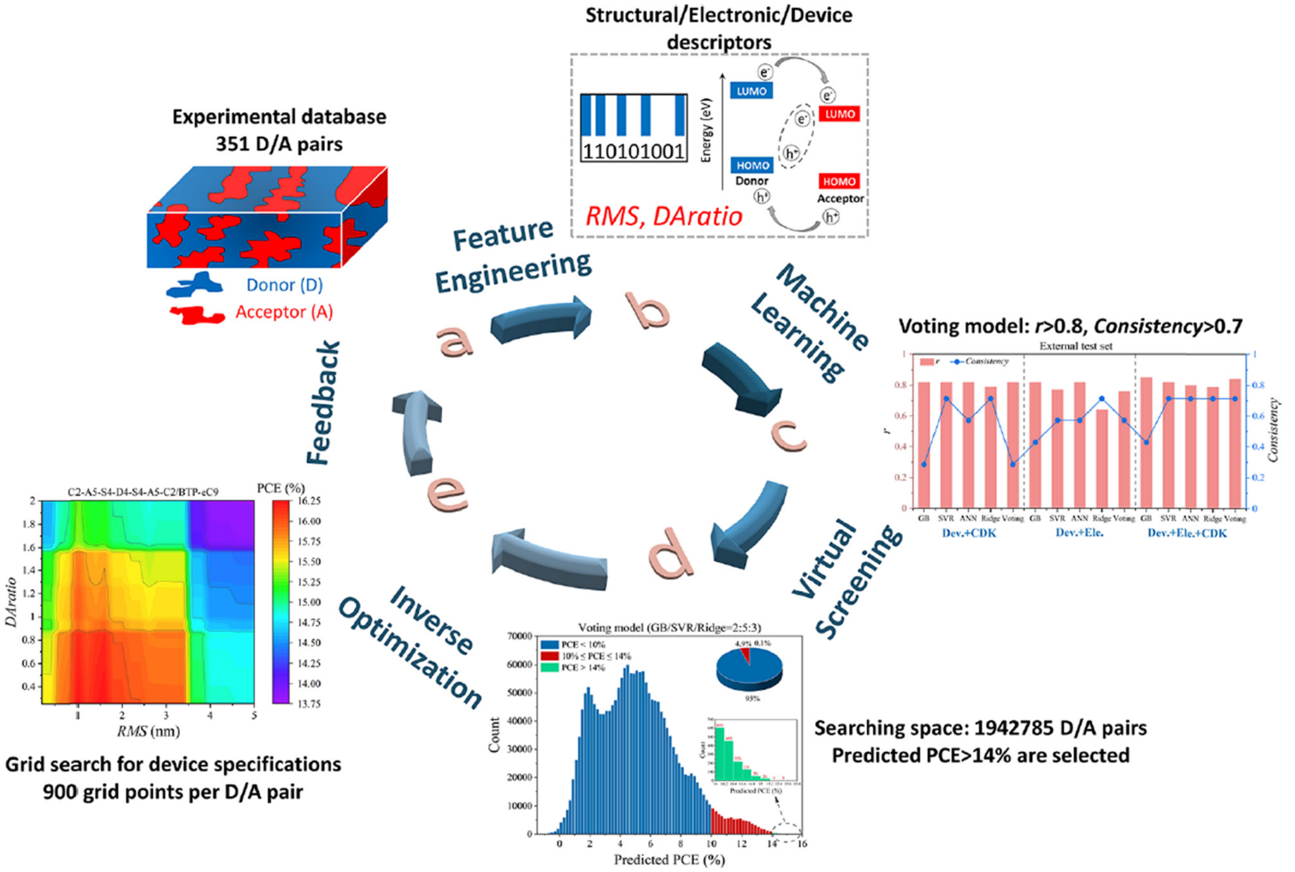

Fig. 13 Workflow of screening and optimization of D/A pairs and device parameters. Reproduced with permission[51] Copyright © 2021 American Chemical Society.

information on coating thickness and material composition ratios at various coating positions indirectly. The RF model was trained on the 2218 dataset, and for the prediction set a high-resolution 3D space based on DD for PM6, Y6, and IT4F was created (80 00 000 rows). Experimental confirmation of the prediction led to a PCE of 10.2% in the expected thickness range for the high efficiency formulation (PM6 : Y6 : IT-4F = 1 : 1.08 : 0.27).

Lu *et al.*[91] created a dataset of 717 D : A pairs of OSCs from the literature (379 papers, published during 2015–2020). In their dataset, donors are both small molecules and polymers (# = 192), while all acceptors are NFAs (# = 377). Input descriptors were calculated using Dragon 7 software[132] (5270 descriptors for each donor and acceptor). As demonstrated in Fig. 14, feature elimination was employed to remove unnecessary features. For the prediction of PCE, four ML algorithms (XGBoost, Decision tree, KNN, and RF) were compared, and XGBoost performed the best ($r = 0.79$). The physical meaning of the significant features was studied by adopting the SHAP approach to better characterize the relationship between key features and PCE. Using high-throughput screening on 76 814 D : A combinations using XGBoost, 10 D : A combinations with high PCE were screened. For ensuring the properties of high PCE screened out D : A pairs, photoelectric properties were calculated using DFT, TDFT, and Marcus charge transfer theory.

Banerji *et al.*[106] in 2022 used the CNN model to predict HOMO/LUMO levels using the HCEP dataset. SMILES strings are converted to 2D RGB images of chemical structures and are used as input for the CNN model. Transfer learning is employed to overcome the poor performance of the deep learning model on a small dataset. The model is initially trained on the HCEP dataset and then fine-tuned (retrained) on the HOPV dataset using a small learning rate. The results are also compared with a use-case dataset (commercially available 26 polymer donors with experimentally measured and DFT estimated FMO) consisting of commercially available donor materials. An illustration of the complete workflow is provided in Fig. 15.

In 2022, Sun *et al.*[133] manually created a dataset of 29 OSCs from the literature employing Y6 and its derivative as an acceptor material and only two donor materials (PBDB-T or PBDB-TF). With LOOCV, the random forest model shows promising results and is further used to screen potential acceptor molecules. Acceptor molecules were split into three parts, end acceptor unit (A1), donor unit (D1), and core acceptor unit (A2), and further encoded with the one-hot-encoding approach. From the dataset all A1–D1–A2 permutations were created as an acceptor search space for potential acceptor molecules with high PCE. With both the donors, the number of D : A combinations were 1296, and PCE predicted by the RF

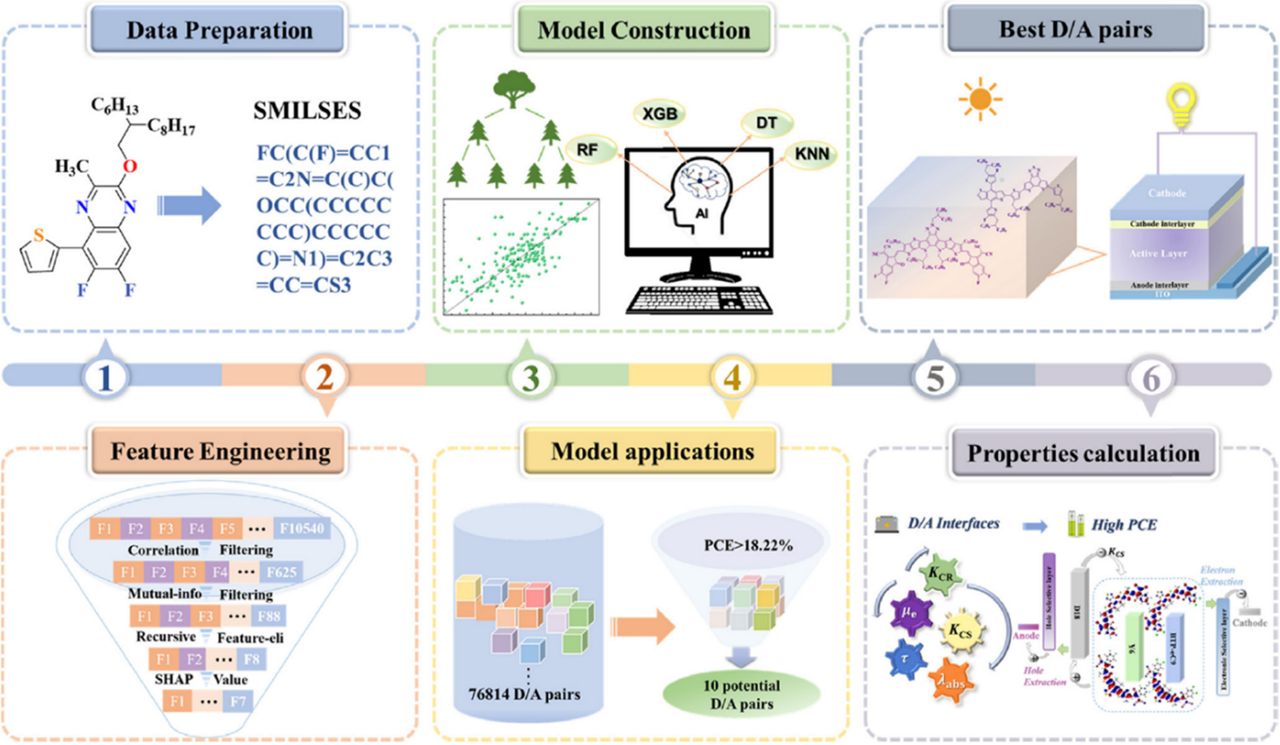

Fig. 14 The flowchart of the strategy integrating machine learning (ML) and density functional theory (DFT) for the screening of promising D/A combinations. Reproduced with permission[91] Copyright © 2022 The Authors. Published by Elsevier Ltd.

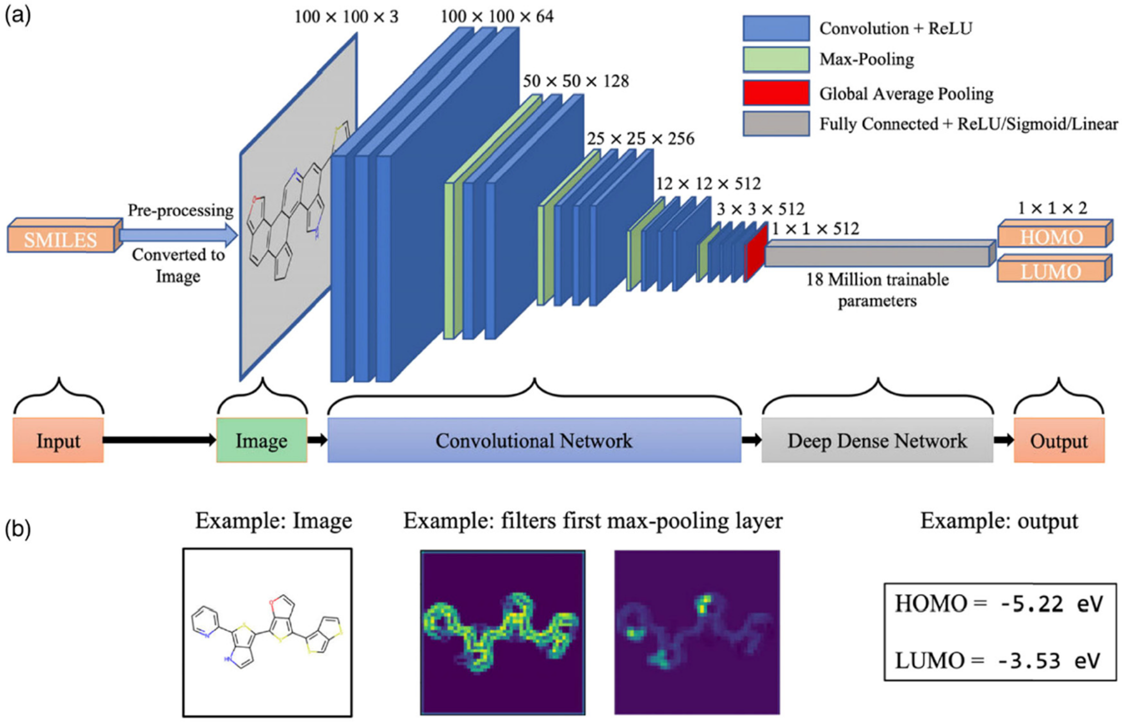

Fig. 15 (a) Illustration of the convolutional neural network, showing the preprocessing step, the convolutional network (with rectified linear unit (ReLU) activation function), max-pooling, global averaging, and fully connected layers (with activation functions). (b) Examples of (left) a molecular image, (middle) the output of two filters after the first max-pooling layer, and (right) the output of the model. Reproduced with permission[106] Copyright © 2022 The Authors. Advanced Theory and Simulations published by Wiley-VCH GmbH.

model was greater than 17% for 25 D : A combinations. The model predicts that molecules with high PCE possess side chains of medium length. Five high-performing acceptors discovered by ML are shown in Fig. 16. The study on five high-performing acceptor molecules with different donors reveals that the difference in the photoelectric properties of these molecules is mainly because of FMO and electrostatic potential.

Hutchison *et al.*[134] discovered novel NFAs by screening 5426 NFAs using a genetic algorithm and various sequences, symmetry and building blocks were analyzed. PBDB-T-SF is used as the donor material for this study because of its strong absorption in UV and high-energy visible regions. Out of 5426 NFAs, 1087 predicted PCE greater than 18%, and 159 predicted PCE greater than 20%. From the study, the A–D–D–D–A or A–A–D–A–A sequence was found to be best for high PCE. Current terminal acceptor units used in common NFAs were determined to be top performers in the GA candidates when picking individual building blocks.

Solubility of active materials is one of the crucial factors governing the PCE of OSCs, and it depends on the length of the solubilizing side alkyl chain. In studies with experimentation validation of ML models, it is found that PCE predictions are overestimated, and the materials are not highly soluble.[82] Unavailability of failure data (failed experiments with low or 0 PCE) can be considered as the major cause of overestimated predictions of ML models.[85] Recently Saeki *et al.*[85] used artificially generated failure data for training the random forest model on NFA based OSCs. To create the failure dataset (# = 875) the authors replaced three 2-ethylhexyl (EH) with methyl (Me) in the previously reported dataset (# = 1295)[83] as shown in Fig. 17(a). D–A polymers with Me group solubility will be low, and the PCE is assumed to be zero. Using the RF model with Mordred descriptors on the experimental + failure dataset achieved $r = 0.84$ as shown in Fig. 17(b), and most of the important features revealed by the model belong to polymer donors (Fig. 17(c)). The model was also tested for new sets of polymers (Fig. 17e and f (# = 19 613) and Fig. 17(g and h) (# = 2 00 932)). The study revealed that for both datasets with inclusion of failure data, the prediction range widens, and predictions of insoluble polymers get shifted to lower PCE, making model predictions more realistic. To check the potential of this model, the authors performed experimental validation by designing and synthesizing 12 new polymers with four types of backbone and different alkyl chains as shown in Fig. 17(i). For each polymer, predicted values with and without failure data are compared with the experimental value as shown in Fig. 17(j). The results reveal that inclusion of failure data gives more realistic predictions and holds potential for discovery of high-performance materials for OSCs.

3.3. Mix based OSCs

The HOPV dataset[92] published in 2016 is a collection of 350 organic small molecules and polymers manually collected from the literature that were used as p-type materials in OSCs. This model has been widely used to train QSPR models. Using the Scharber model, J_{sc}, V_{oc}, and PCE were calculated, while FMOs and bandgaps were calculated using the DFT functionals BP86, B3LYP, M06-2x, and PBE0. Many research papers have used this dataset.[54,93–95]

Aspuru Guzik *et al.*[93] (2016) calculated J_{sc}, V_{oc}, and PCE using the Scharber model. The authors calibrated quantum chemical calculations using the Gaussian approach to improvise the results. The study was conducted with the HOPV dataset on HOMO, LUMO, bandgap, PCE, V_{oc}, and J_{sc}. The results in

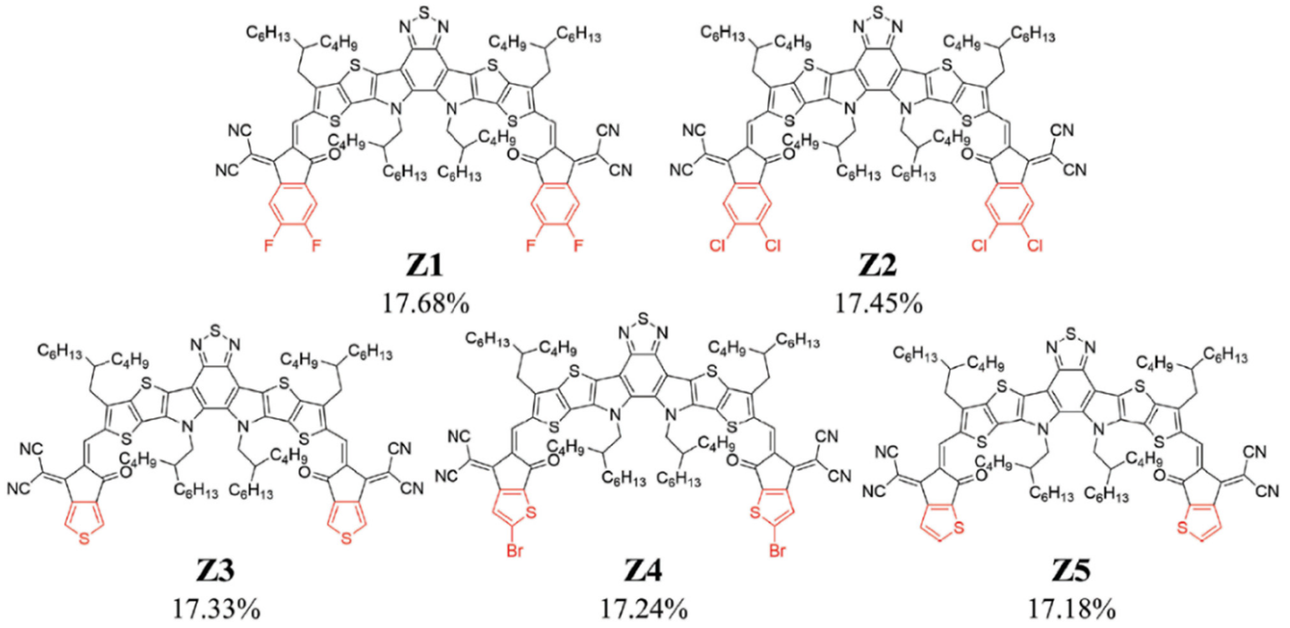

Fig. 16 Five typical machine learning predicted high-performance acceptor molecules with different acceptor units at the end groups (highlighted in red) and their predicted PCE. Reproduced with permission[133] Copyright© 2022 The Authors. Advanced Science published by Wiley-VCH GmbH.

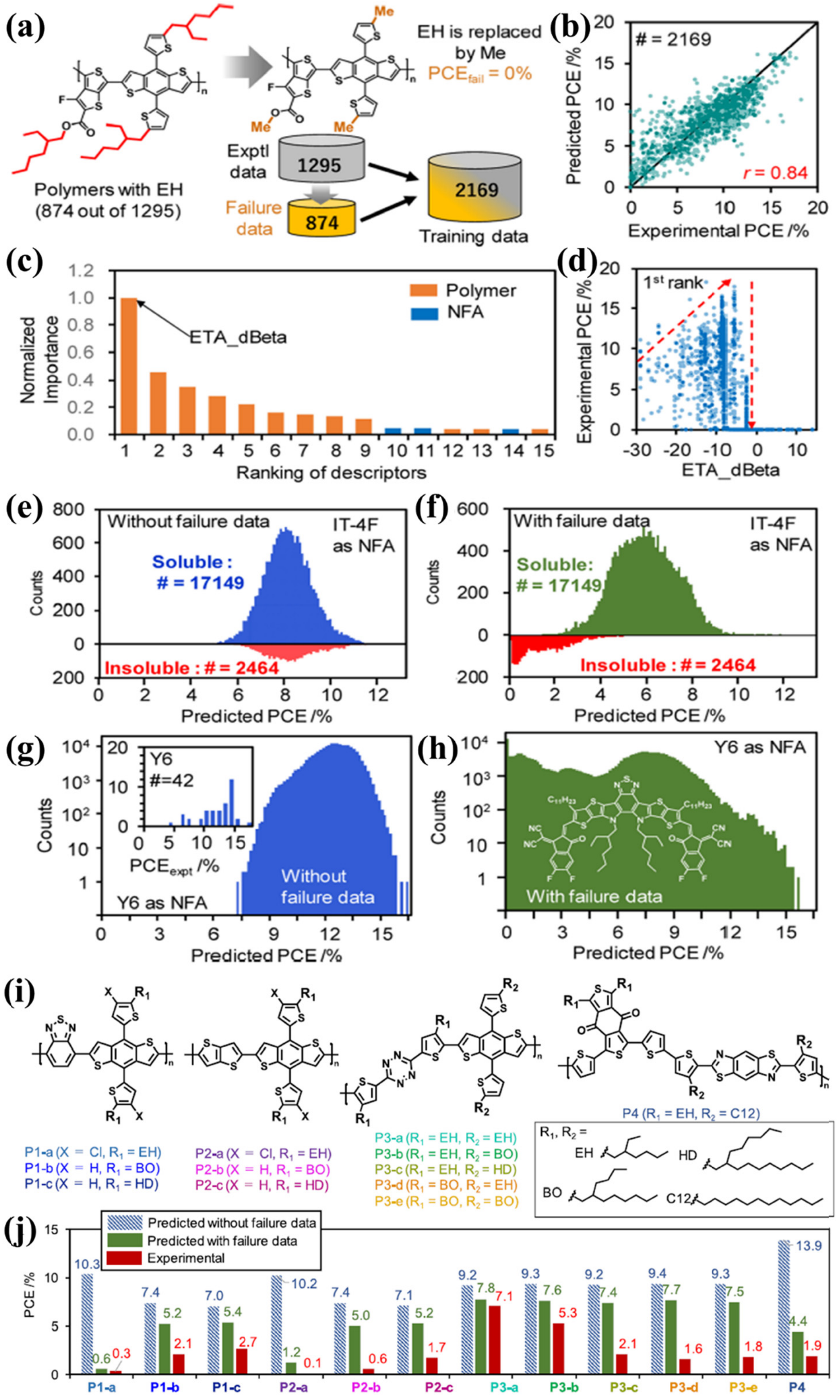

Fig. 17 (a) Procedure of generating failure data of polymers. (b) Correlation of PCE$_{pred}$ *vs.* PCE$_{exptl}$ calculated by RF CV using failure data. The *r* and # values are appended. (c) Ranking of normalized importance of the RF model. The orange and blue bars are polymer and NFA-related parameters, respectively. (d) PCE$_{exptl}$ *vs.* ETA_dBeta, the number one-ranked parameter by importance in the RF model. The red arrows represent a trend. (e and f) PCE$_{pred}$ distribution of 19 613 virtually generated polymers using IT-4F as the NFA. The PCE$_{pred}$ was predicted by the RF (e) without and (f) with failure data. The categories of soluble/insoluble polymers were labeled by another RF model. (g and h) PCE$_{pred}$ distribution of 200 932 virtually generated polymers using Y6 as the NFA. The PCE$_{pred}$ was predicted by the RF (g) without and (h) with failure data. The inset in (g) is the PCE$_{exptl}$ histogram for the dataset. The chemical structure of Y6 is superimposed on (h). (i) Chemical structures of the synthesized polymers. (j) Predicted PCE of twelve polymers (P1–P3 assume IT-4F as the NFA; P4 assumes Y6 as the NFA) calculated by the RF model without failure data (meshed blue) and with failure data (solid green). Reproduced with permission[85] Copyright © 2022 American Chemical Society.

terms of Pearson's r were improved after the calibration was applied.

Balasubramanian *et al.*[54] generated novel polymer donors for OSCs using the transfer learning scheme based on long short-term memory (LSTM). Since the SMILES notation is a chemical language, LSTM can understand this language's grammar and vocabulary. Novel polymer donor SMILES strings were generated with the LSTM model established on a large dataset ($\sim$1 million) from the GDB17 chemical database[135] and were further tuned on a smaller ($\sim$1400 conjugated molecules) dataset.[66,92] A total of 1000 molecules were generated, out of which 90% were valid SMILES strings with an average Tanimoto similarity of 0.42 ± 0.010. The authors also created a response surface plot for PCE using PCA and found that newly generated polymer donors lie in the region of high PCE.

Russo *et al.*[95] demonstrated the pre-screening potential of machine learning in accurately predicting OSCs' material properties and bypassing the need for expensive DFT calculations. For the HOPV dataset, the authors have used molecular signature descriptors (chemical fragments of donors and acceptors) and one-hot descriptors to make their model interpretable. The model could predict DFT calculated PCE with a standard error of ±0.5.

Sun *et al.*[96] in 2019 gathered a dataset of 1719 donor materials (mix of polymers and small molecules) from the literature with fullerene or non-fullerene acceptors. To build structure–property relationships, ASCII stings, fingerprints, descriptors, and images were used as input for the ML model. For the classification study, two classes were created, low performance (PCE in the range 0–2.99) and high performance (PCE higher than 3%). The results reveal that fingerprints with length over 1000 bits perform much better than other descriptors. 81.76% accuracy was obtained in predicting PCE by the RF model using fingerprints as input. The authors also verified the potential of the ML model by synthesizing 10 new donor materials for OSCs, and the predicted PCE values were in good agreement with the experimental values.

Troisi *et al.*[97] studied the optimization of multicomponent materials for BHJ OSCs. For this purpose, a dataset of 320 D : A combinations (262 unique donors and 76 unique acceptors) was manually collected from the literature, and PCE was predicted with different ML models (KRR, SVR, k-NN, and Gaussian process regressor (GPR)). KRR with LOOCV gave the best results (r = 0.78). Model performance was also verified on recently published D : A combinations and impressive results were obtained.

Troisi *et al.*[46] in 2020 created a dataset of 566 D : A pairs (513 distinct donors and 33 distinct acceptors) and studied the effect of increasing the descriptor set on prediction of PCE using K-nearest neighbours (KNN), kernel ridge regressor (KRR), and support vector regression (SVR). The authors divided the descriptors used in this study into five groups: (1) molecular topology, (2) properties related to molecular size, (3) properties related to molecular energy levels, (4) absorption properties, and (5) mixing properties. Fingerprints were computed using the RDKit package,[118] and miscibility properties were calculated using SwiddADME.[136] The authors concluded that excited state and miscibility-related properties do not improve the model's performance since the information they carry is already encoded within the structural fingerprint. It is important to remember that not all machine learning models have the exact computing cost. Some low-cost methods, such as simple physical descriptors and structural information, already yield good results.

Usually, ML models perform well on the class of compounds already used in the training set. However, they do not perform well on a new class of compounds (extrapolation) that have not been seen in the training stage. Troisi *et al.*[101] in 2020 used leave-one-group-out (LOGO) cross-validation to predict the PCE of NFA materials from completely new families and accelerate materials discovery. The dataset used in this study consists of 566 D : A pairs. The authors have shown that LOGO significantly improves PCE prediction of unseen materials above and below the median efficiency. The authors also concluded that physical descriptors provide much more reliable results than fingerprints when extrapolating to new chemical families.

Tandem OSCs can simultaneously address narrow absorption window and thermalization problems by stacking two or more cells with a complementary absorption range.[137] Lee[102] created a dataset of 70 tandem OSCs (37 are conventional and 33 are inverted) to predict PCE with RF and SVR. FMOs for both cells (bottom sub-cell near ITO and top sub-cell near the metal electrode) were used as input descriptors. The RF model performed the best (R^2 = 0.69) on the test set. Using feature importance, FMO of the donor (bottom sub-cell near ITO) was found to be the most crucial feature in predicting the PCE of tandem OSCs. The best results were achieved by RF followed by XGBoost. Feature importance by the RF model suggests high importance to LUMO of acceptor, while relatively low importance is assigned to HMO and LUMO of the third component. The authors also performed a two-class classification study with class1 (PCE < 16%) and class2 (PCE > 16%). Again the RF model performed the best with an accuracy score of 0.97 on the test set.

Kettle *et al.*[99] in 2020 used a dataset of 1850 OSCs (2011–2019) with corresponding device characteristics, performance, and stability data. The stability and performance of OSCs are estimated using the sequential minimal optimization regression (SMOreg) model. The dataset is acquired from their earlier work.[98] The key attributes related to OSC degradation were identified using the initial efficiency (Eo) and time taken to reach 80% of the initial value (T80). The dataset was separated for tests conducted under ISOS-L and ISOS-D. The choice of light spectrum and active materials were found to be the key parameters to increase stability for ISOS-L testing, while material-depending attributes and encapsulation were found to be the key parameters for ISOS-D testing.

In 2020, Brabec *et al.*[56] introduced robotization of the lab for OSC studies. With their novel robotized film creation technology, 6048 films can be produced per day. This automated experimentation platform is integrated with an active learning technique (Bayesian optimization) for 4D parameter space (PTB7-Th or PBQ-QF, P3HT, oIDTBR, PCBM) of quaternary

OSC blends. High-throughput experimentation is done by using the ChemOS software package. The study demonstrates lower stability for PTB7-Th rich blends than P3HT and PBQ-QF rich blends. Moreover, the active learning technique is able to find stability maxima and minima with a 93% reduction of sample. This study demonstrates that with robotization, 2000 combinations can be screened with less than 10 mg material, and complex active layer parameter optimization problems can be solved. The automated platform and workflow of high throughput experimentation are represented in Fig. 18.

By commonly used fingerprints like MACCS[116] and Pubchem,[139] we cannot correlate each bit of fingerprint with a specific substructure. To overcome this limitation of fingerprints, Sun *et al.*[103] designed the La FREMD fingerprint for expressing 6180 different fragments (bits). Integration of the La FREMD fingerprint and ML allows rapid design and screening of donor materials for high-efficiency OSCs. This framework is used on a dataset of 1758 experimentally reported donor materials. Using the RF model, 15 most important fragments were identified for high PCE and were compared based on the newly described variable frequency difference (FD). As shown in Fig. 19, after identifying important building blocks, a library of 18 960 small molecule donors is created with their combination. For rapid screening from this library, four ML models (ANN, GBRT, RF, and SVR) were trained on the donor material database (1758) with daylight fingerprint as the input descriptor. The GBRT model gave the lowest RMSE (2.05) and 6337 molecules predicted PCE greater than 8%. 20 promising molecules were selected based on molecular symmetry with Y6 as the acceptor. For this purpose a new dataset of 44 OSCs with the Y6 or Y6 derivative was used to train the GBRT model for PCE prediction and test it on 20 promising molecules selected earlier. 5 out of 20 promising molecules predicted PCE greater than 15%.

Hutchinson *et al.*[104] created a model for NFA based OSCs using 47 input descriptors to predict FF, J_{sc}, V_{oc}, and PCE. To speed up the calculation, they used sTD-DFT[140] which is 2–3 times faster than TD-DFT. A dataset of 84 D : A combinations was used for modeling, out of which 6 are polymer donors and 66 are NFAs and 2 are FAs. The results reveal that the sTD-DFT based model can predict PCE with RMSE = 1.60 ± 0.04%.

A methodology is required to rapidly identify optimal device configurations that optimize net energy output while minimizing the environmental impact of OPVs. Kettle *et al.*[100] created a dataset of 1580 OSC devices (2011–2020) with reported PCE and stability data. The objective is to find optimal OSC materials and device architectures that are environmentally friendly. For each device, the net energy (E_{Net}) is also calculated, which is a function of time taken to degrade to 80% of its stabilized initial output power (T_{80}), time taken to drop to 80% of initial power at t = 0 (T_{S80}), and embodied energy (E_{Emb}). The SMOreg model is used to analyze E_{Net} by using structural components of each device. GA clustering was used to determine the optimal material sets for OPV ENet output. This study uses the

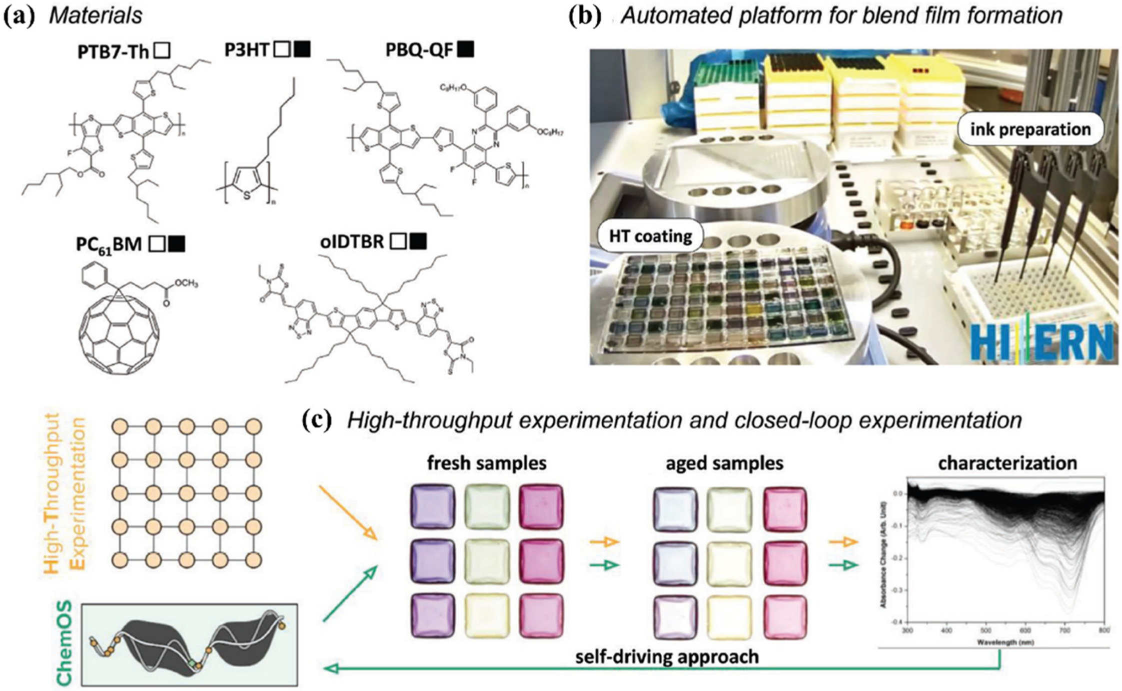
(a) Materials
PTB7-Th □
P3HT □■
PBQ-QF ■
PC61BM □■
oIDTBR □■
(b) Automated platform for blend film formation
ink preparation
HT coating
HI ERN
(c) High-throughput experimentation and closed-loop experimentation
High-Throughput Experimentation
ChemOS
fresh samples
aged samples
characterization
Absorbance Change (Arb. Unit)
Wavelength (nm)
self-driving approach

Fig. 18 (a) Representations of the three polymer donors and the two small molecule acceptors. Note that the first quaternary system consists of P3HT, PBQ-QF, PCBM, and oIDTBr (■) and the second consists of P3HT, PTB7-Th, PCBM, and oIDTBr (□). (b) Side view of the automated platform for ink formulation, coating, and characterization. (c) Experimental workflow with the two approaches adopted in this study: conventional high-throughput experimentation *via* grid and the self-driving approach with the ChemOS software package. Reproduced with permission[138] Copyright, © 2020 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

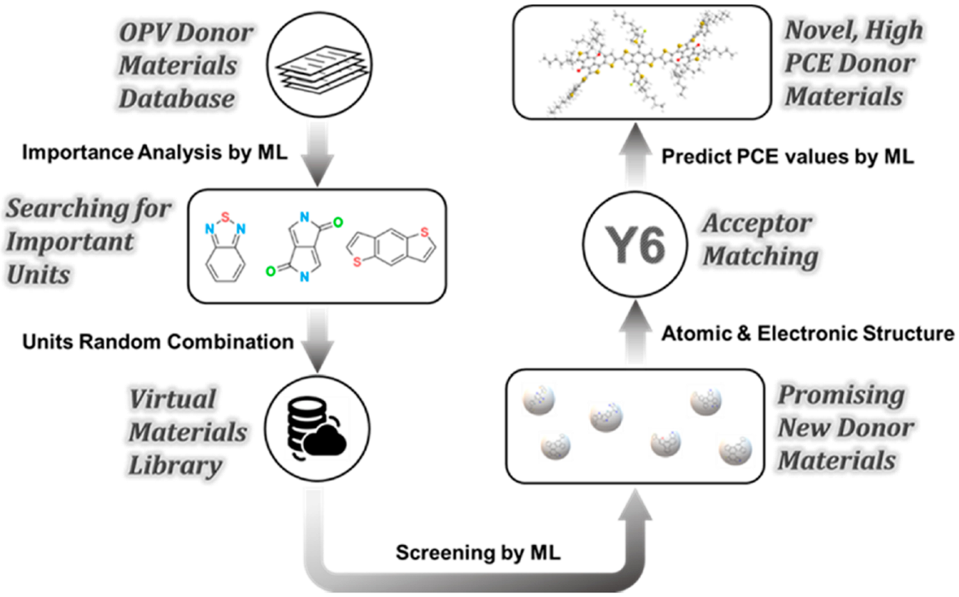
OPV Donor Materials Database
Importance Analysis by ML
Searching for Important Units
Units Random Combination
Virtual Materials Library
Screening by ML
Promising New Donor Materials
Atomic & Electronic Structure
Y6
Acceptor Matching
Predict PCE values by ML
Novel, High PCE Donor Materials

Fig. 19 Scheme of the AI design framework for developing high-performance OPV donor materials. Reproduced with permission[103] Copyright © 2021 American Chemical Society.

"GenClust++" clustering method. By using the genetic search method clustering, eco-friendly device configurations are obtained. The complete workflow is depicted in Fig. 20. The active layer materials DRCN7T, DR3TSBDT, ZnPc, PDPP4T-2F,

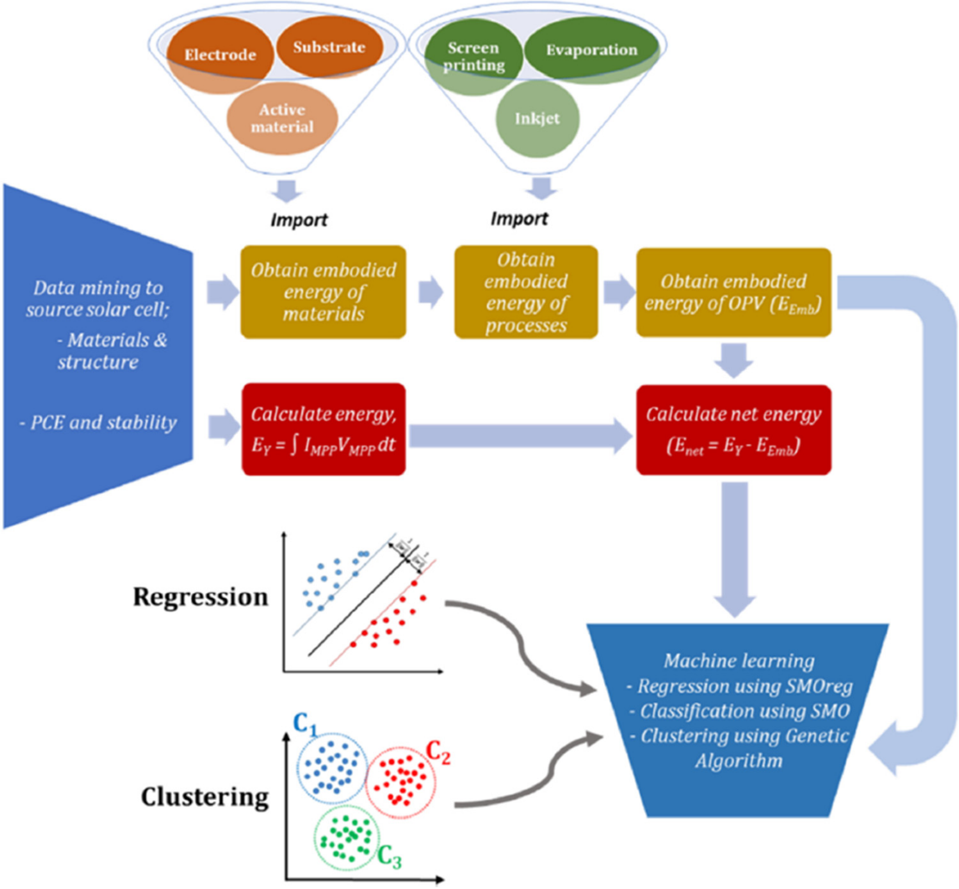
Electrode
Substrate
Active material
Screen printing
Evaporation
Inkjet
Import
Import
Data mining to source solar cell;
- Materials & structure
- PCE and stability
Obtain embodied energy of materials
Obtain embodied energy of processes
Obtain embodied energy of OPV (E_{Emb})
Calculate energy, $E_Y = \int I_{MPP}V_{MPP}dt$
Calculate net energy ($E_{net} = E_Y - E_{Emb}$)
Regression
Clustering
C_1
C_2
C_3
Machine learning
- Regression using SMOreg
- Classification using SMO
- Clustering using Genetic Algorithm

Fig. 20 Schematic of how material and energy costs are acquired and stages of analysis using ML GAs. Reproduced with permission[100] Copyright © 2022 American Chemical Society.

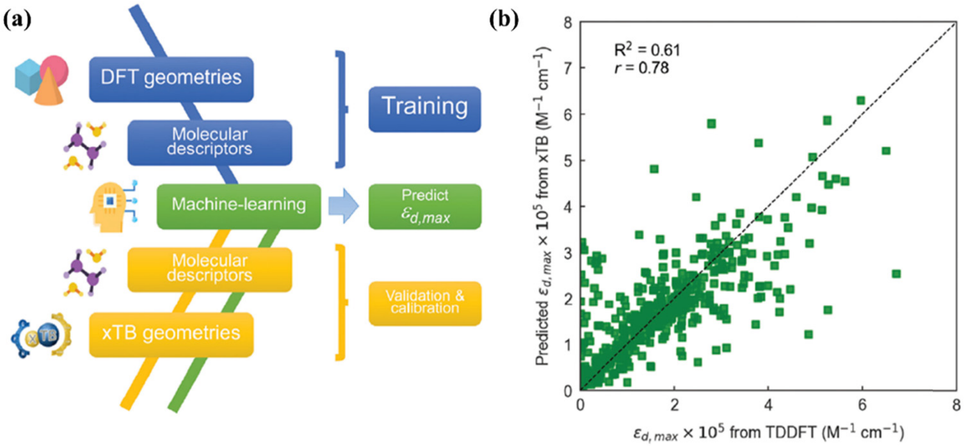

Fig. 21 (a) ML workflow used in this work to draw ed,max predictions. A RF model is trained on TDDFT data and interpolated (validated) on xTB geometries, including also their corresponding molecular descriptors. To improve the accuracy of the model, energy levels obtained using the GFN2-xTB Hamiltonian require calibration with TDDFT values. (b) Leave-one-out interpolation of the resulting RF model using three input molecular descriptors (including calibrated energy levels) and a 64-bit Morgan fingerprint vector. Reproduced with permission[141] Copyright, 2022, Royal Society of Chemistry (RSC).

PCDTBT (donors), and IT-4F, C61 (acceptors) show promise for delivering positive net energy. The authors concluded that trained and validated models could accurately predict the efficiency, stability, and embodied energy of an OPV.

OPV performance can be improved in a variety of ways by better understanding how smart molecular design can increase light harvesting on its own. The device's macroscopic short-circuit current density is closely tied to light absorption, which is the first stage in generating an electrical charge. High absorption should, in principle, lead to strong emission due to the reciprocity relationship between absorption and emission, lowering nonradiative energy losses and benefiting open-circuit voltage. For identifying structure–absorption relationship, Nelson *et al.*[141] investigated 500 unique organic molecules with DFT and TDDFT to study absorption strength and unravel structural features that lead to high absorption strength. The authors found that the absorption strength calculated experimentally and by TDDFT is in good agreement for NFAs and fullerene and suggested the use of TDDFT calculations for further modeling. From DFT optimized geometries, 6000 molecular descriptors were created, and the highest correlation with absorption strength was given by $\varepsilon_{d,max}$ (experimentally measured maximum molar extinction coefficient), $\lambda_{1,p}$ (size of the molecule in the direction of maximum atomic polarizability), and C2SP2 (number of sp^2 hybridized carbon atoms bound with two other carbons). TDDFT and ML studies found that molecular linearity, planarity, polarizability and number of Π-conjugated carbon atoms also correlate strongly with absorption strength. Moreover, the authors created RF model using LOOCV to predict ed, max by using semiempirical extended tight-binding (xTB) Hamiltonians which is 3000 times faster than using DFT. The workflow of this study along with prediction results is represented in Fig. 21.

4. Future scope and challenges

With the growing interest in ML applications for materials science, significant research is also conducted on OSCs using ML on available experimental data. With continuous development in algorithms and enhancement of computational power, we may anticipate the trend of ML-OSC research to continue.

As the PCE has achieved acceptable levels in recent years, the focus has switched from efficiency improvement to stability, which remains an obstacle to the widespread use of the technology. There is still a lot we do not know about how environmental stress variables like temperature, humidity, and oxygen affect the degrading behavior of materials and, by extension, the durability of solar cells under the conditions they will be used or stored.[142]

Various problems faced by researchers in this direction are discussed here.

4.1. Selection of a suitable training set

The same material is being used in the literature with different names. The solution is to use the canonical SMILES string after gathering the dataset.

4.2. Dataset size

OSC datasets are pretty small, and k-fold cross-validation gives poor results. The solution is to use leave-one-out cross-validation for small datasets.

4.3. Extrapolation

For OSCs, the available dataset for training ML models is small and chemical structures are also large and complex. Thus, ML models often struggle in extrapolation tasks and give poor results on the dataset that is never seen during the training. The solution is to use a scaffold-based train-test split or leave one group out cross-validation.[101] If large datasets are available, feature engineering can improve the extrapolation ability of ML models.

4.4. Simple input descriptors

To quickly make predictions for previously untested materials or combinations after saving a machine learning model, it is important to have quick access to the descriptors used in the model. Input descriptors should be easy to calculate and should be sensitive to the target variable. For example, simple input descriptors such as 'Number of Nitrogen and Oxygen' or 'Number of valence electrons the molecule has' could be easily obtained using RDKit. The problem can be solved by employing descriptors that are computed immediately from a chemical structure or SMILES string.

4.5. Inclusion of failure data

During training, the model learns from training data, and usually failure data are not reported in publications and hence are not included in experimental datasets. Because of unavailability of the failure dataset, models' predictions are usually overestimated and highly deviated. With the inclusion of the failure dataset, the model can understand device physics in a much better way and avoid overestimated prediction.

4.6. Active learning techniques

Active learning techniques can guide OSC experimentalists to find most suitable active materials from vast search space in a few number of experiments. Active learning could be also used for finding the most optimal device fabrication parameters such as ratio of donor and acceptor in the BHJ active layer, thickness of the active layer, annealing temperature of the active layer, *etc.*

4.7. Experimental validation

To verify the results predicted by ML models, experimental validation is required. Various studies have demonstrated that ML models predict photovoltaic properties on the test set with high accuracy. But when predicted on novel materials, the results deviate by a considerable margin. The reason for this is that a number of experimental parameters such as D : A ratio, spin speed, solvent, processing temperature, *etc.*, are not usually taken as input while training.

The research of ML in the field of OSCs is still in its early stage. While many models and/or algorithms have been developed, very few new materials, processes or structures are demonstrated successful. With the increasing volume of data accumulated and research experience, the accuracy of models is anticipated to be improved substantially. We foresee that AI can truly aid scientists in gathering knowledge of various domains and to develop promising materials and experimental parameters for OSCs in the near future.

5. Conclusion and perspectives

Tremendous efforts have been devoted to developing ML models, which can learn from the existing performance data of OSCs, to help the researchers to predict and explore the properties of new organic materials and devices. The accuracy from the ML prediction has been improved substantially in many respects of OSCs, including efficiency and stability. ML models now can also assist and simplify the optimization of experimental parameters for device fabrication.

On the other hand, as the data size of OSCs is relatively small, it is very critical to develop ML algorithms and/or design workflows that can be trained from a small-sized dataset. Further, new descriptors and structural fingerprints can improve models' accuracy and help develop new materials. Although many reports of ML models demonstrate extrapolation ability, accurate predictions are still very challenging. Such prediction ability is more important because the results can indeed save many research resources and accelerate the development progress. Most current "screening" functions of the ML models can only handle chemical structures built from known blocks and units. Finally, the automation from material design, synthesis work, experimental data collection to result analysis is worthy of attention. With a complete research automated loop, we can explore much larger materials space in a limited time. The introduction of ML models and AI to the development of OSCs is still in its infancy. We witnessed the beginning of this field, and only very limited materials are developed using the ML models currently. We anticipate that more organic materials will be synthesized, and the performance of OSCs can eventually be pushed to the limit with the assistance from ML models.

Conflicts of interest

There are no conflicts to declare.

References

1 O. Almora, D. Baran, G. C. Bazan, C. Berger, C. I. Cabrera, K. R. Catchpole, S. Erten-Ela, F. Guo, J. Hauch, A. W. Y. Ho-Baillie, T. J. Jacobsson, R. A. J. Janssen, T. Kirchartz, N. Kopidakis, Y. Li, M. A. Loi, R. R. Lunt, X. Mathew, M. D. McGehee, J. Min, D. B. Mitzi, M. K. Nazeeruddin, J. Nelson, A. F. Nogueira, U. W. Paetzold, N. Park, B. P. Rand, U. Rau, H. J. Snaith, E. Unger, L. Vaillant-Roca, H. Yip and C. J. Brabec, Device Performance of Emerging Photovoltaic Materials (Version 2), *Adv. Energy Mater.*, 2021, **11**, 2102526.

2 A. Karki, A. J. Gillett, R. H. Friend and T. Nguyen, The Path to 20% Power Conversion Efficiencies in Nonfullerene Acceptor Organic Solar Cells, *Adv. Energy Mater.*, 2021, **11**, 2003441.

3 L. Duan and A. Uddin, Progress in Stability of Organic Solar Cells, *Adv. Sci.*, 2020, **7**, 1903259.

4 J. Wu, M. Gao, Y. Chai, P. Liu, B. Zhang, J. Liu and L. Ye, Towards a bright future: The versatile applications of organic solar cells, *Mater. Rep. Energy*, 2021, **1**, 100062.

5 K. Khandelwal, S. Biswas, A. Mishra and G. D. Sharma, Semitransparent organic solar cells: From molecular design to structure-performance relationships, *J. Mater. Chem. C*, 2022, **10**, 13–43.

6 S. Lee, D. Jeong, C. Kim, C. Lee, H. Kang, H. Y. Woo and B. J. Kim, Eco-friendly polymer solar cells: Advances in

green-solvent processing and material design, *ACS Nano*, 2020, **14**, 14493–14527.

7 Y. Cui, L. Hong and J. Hou, Organic Photovoltaic Cells for Indoor Applications: Opportunities and Challenges, *ACS Appl. Mater. Interfaces*, 2020, **12**, 38815–38828.

8 E. Ravishankar, R. E. Booth, C. Saravitz, H. Sederoff, H. W. Ade and B. T. O'Connor, Achieving Net Zero Energy Greenhouses by Integrating Semitransparent Organic Solar Cells, *Joule*, 2020, **4**, 490–506.

9 J. Hou, O. Inganas, R. H. Friend and F. Gao, Organic solar cells based on non-fullerene acceptors, *Nat. Mater.*, 2018, **17**, 119–128.

10 A. Wadsworth, M. Moser, A. Marks, M. S. Little, N. Gasparini, C. J. Brabec, D. Baran and I. McCulloch, Critical review of the molecular design progress in non-fullerene electron acceptors towards commercially viable organic solar cells, *Chem. Soc. Rev.*, 2019, **48**, 1596–1625.

11 F. Zhao, H. Zhang, R. Zhang, J. Yuan, D. He, Y. Zou and F. Gao, Emerging Approaches in Enhancing the Efficiency and Stability in Non-Fullerene Organic Solar Cells, *Adv. Energy Mater.*, 2020, **10**, 2002746.

12 A. Armin, W. Li, O. J. Sandberg, Z. Xiao, L. Ding, J. Nelson, D. Neher, K. Vandewal, S. Shoaee, T. Wang, H. Ade, T. Heumüller, C. Brabec and P. Meredith, A History and Perspective of Non-Fullerene Electron Acceptors for Organic Solar Cells, *Adv. Energy Mater.*, 2021, **11**, 1–42.

13 D. Luo, W. Jang, D. D. Babu, M. S. Kim, D. H. Wang and A. K. K. Kyaw, Recent progress in organic solar cells based on non-fullerene acceptors: materials to devices, *J. Mater. Chem. A*, 2022, **10**, 3255–3295.

14 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 and Y. Zou, Single-Junction Organic Solar Cell with over 15% Efficiency Using Fused-Ring Acceptor with Electron-Deficient Core, *Joule*, 2019, **3**, 1140–1151.

15 Q. Wei, W. Liu, M. Leclerc, J. Yuan, H. Chen and Y. Zou, A-DA′D-A non-fullerene acceptors for high-performance organic solar cells, *Sci. China: Chem.*, 2020, **63**, 1352–1366.

16 B. Lu, J. Wang, Z. Zhang, J. Wang, X. Yuan, Y. Ding, Y. Wang and Y. Yao, Recent progress of Y-series electron acceptors for organic solar cells, *Nano Sel.*, 2021, **2**, 2029–2039.

17 Y. Cui, Y. Xu, H. Yao, P. Bi, L. Hong, J. Zhang, Y. Zu, T. Zhang, J. Qin, J. Ren, Z. Chen, C. He, X. Hao, Z. Wei and J. Hou, Single-Junction Organic Photovoltaic Cell with 19% Efficiency, *Adv. Mater.*, 2021, **33**, 2102420.

18 L. Zhu, M. Zhang, J. Xu, C. Li, J. Yan, G. Zhou, W. Zhong, T. Hao, J. Song, X. Xue, Z. Zhou, R. Zeng, H. Zhu, C. C. Chen, R. C. I. MacKenzie, Y. Zou, J. Nelson, Y. Zhang, Y. Sun and F. Liu, Single-junction organic solar cells with over 19% efficiency enabled by a refined double-fibril network morphology, *Nat. Mater.*, 2022, **21**, 656–663.

19 C. Li, J. Zhou, J. Song, J. Xu, H. Zhang, X. Zhang, J. Guo, L. Zhu, D. Wei, G. Han, J. Min, Y. Zhang, Z. Xie, Y. Yi, H. Yan, F. Gao, F. Liu and Y. Sun, Non-fullerene acceptors with branched side chains and improved molecular packing to exceed 18% efficiency in organic solar cells, *Nat. Energy*, 2021, **6**, 605–613.

20 Y. Wei, Z. Chen, G. Lu, N. Yu, C. Li, J. Gao, X. Gu, X. Hao, G. Lu, Z. Tang, J. Zhang, Z. Wei, X. Zhang and H. Huang, Binary Organic Solar Cells Breaking 19% via Manipulating Vertical Component Distribution, *Adv. Mater.*, 2022, **34**, 2204718.

21 W. Feng, S. Wu, H. Chen, L. Meng, F. Huang, H. Liang, J. Zhang, Z. Wei, X. Wan, C. Li, Z. Yao and Y. Chen, Tuning Morphology of Active Layer by using a Wide Bandgap Oligomer-Like Donor Enables Organic Solar Cells with Over 18% Efficiency, *Adv. Energy Mater.*, 2022, **12**, 2104060.

22 C. He, Y. Pan, Y. Ouyang, Q. Shen, Y. Gao, K. Yan, J. Fang, Y. Chen, C.-Q. Ma, J. Min, C. Zhang, L. Zuo and H. Chen, Manipulating the D:A interfacial energetics and intermolecular packing for 19.2% efficiency organic photovoltaics, *Energy Environ. Sci.*, 2022, **15**, 2537–2544.

23 L. Zhan, S. Li, Y. Li, R. Sun, J. Min, Z. Bi, W. Ma, Z. Chen, G. Zhou, H. Zhu, M. Shi, L. Zuo and H. Chen, Desired open-circuit voltage increase enables efficiencies approaching 19% in symmetric-asymmetric molecule ternary organic photovoltaics, *Joule*, 2022, **6**, 662–675.

24 Z. Zheng, J. Wang, P. Bi, J. Ren, Y. Wang, Y. Yang, X. Liu, S. Zhang and J. Hou, Tandem Organic Solar Cell with 20.2% Efficiency, *Joule*, 2022, **6**, 171–184.

25 Y. Lin, Q. He, F. Zhao, L. Huo, J. Mai, X. Lu, C. J. Su, T. Li, J. Wang, J. Zhu, Y. Sun, C. Wang and X. Zhan, A Facile Planar Fused-Ring Electron Acceptor for As-Cast Polymer Solar Cells with 8.71% Efficiency, *J. Am. Chem. Soc.*, 2016, **138**, 2973–2976.

26 Y. Lin, F. Zhao, Q. He, L. Huo, Y. Wu, T. C. Parker, W. Ma, Y. Sun, C. Wang, D. Zhu, A. J. Heeger, S. R. Marder and X. Zhan, High-Performance Electron Acceptor with Thienyl Side Chains for Organic Photovoltaics, *J. Am. Chem. Soc.*, 2016, **138**, 4955–4961.

27 S. Dai, F. Zhao, Q. Zhang, T. K. Lau, T. Li, K. Liu, Q. Ling, C. Wang, X. Lu, W. You and X. Zhan, Fused nonacyclic electron acceptors for efficient polymer solar cells, *J. Am. Chem. Soc.*, 2017, **139**, 1336–1343.

28 J. Wang and X. Zhan, Fused-Ring Electron Acceptors for Photovoltaics and beyond, *Acc. Chem. Res.*, 2021, **54**, 132–143.

29 A. Harillo-Baños, X. Rodríguez-Martínez and M. Campoy-Quiles, Efficient Exploration of the Composition Space in Ternary Organic Solar Cells by Combining High-Throughput Material Libraries and Hyperspectral Imaging, *Adv. Energy Mater.*, 2020, **10**, 1902417.

30 X. Rodríguez-Martínez, S. Sevim, X. Xu, C. Franco, P. Pamies-Puig, L. Córcoles-Guija, R. Rodriguez-Trujillo, F. J. del Campo, D. Rodriguez San Miguel, A. J. deMello, S. Pané, D. B. Amabilino, O. Inganäs, J. Puigmartí-Luis and M. Campoy-Quiles, Microfluidic-Assisted Blade Coating of Compositional Libraries for Combinatorial Applications: The Case of Organic Photovoltaics, *Adv. Energy Mater.*, 2020, **10**, 2001308.

31 R. Po, A. Bernardi, A. Calabrese, C. Carbonera, G. Corso and A. Pellegrino, From lab to fab: how must the polymer

solar cell materials design change? – an industrial perspective, *Energy Environ. Sci.*, 2014, **7**, 925–943.

32 J. E. Carlé, M. Helgesen, O. Hagemann, M. Hösel, I. M. Heckler, E. Bundgaard, S. A. Gevorgyan, R. R. Søndergaard, M. Jørgensen, R. García-Valverde, S. Chaouki-Almagro, J. A. Villarejo and F. C. Krebs, Overcoming the Scaling Lag for Polymer Solar Cells, *Joule*, 2017, **1**, 274–289.

33 P. Meredith and A. Armin, Scaling of next generation solution processed organic and perovskite solar cells, *Nat. Commun.*, 2018, **9**, 5261.

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

35 Y. T. Fu, C. Risko and J. L. Brédas, Intermixing at the Pentacene-Fullerene Bilayer Interface: A Molecular Dynamics Study, *Adv. Mater.*, 2013, **25**, 878–882.

36 Y. Shin, J. Liu, J. J. Quigley, H. Luo and X. Lin, Combinatorial design of copolymer donor materials for bulk heterojunction solar cells, *ACS Nano*, 2014, **8**, 6089–6096.

37 Y. Imamura, M. Suganuma and M. Hada, Computational Study on the Search for Non-Fullerene Acceptors, Examination of Interface Geometry, and Investigation of Electron Transfer, *J. Phys. Chem. C*, 2019, **123**, 17678–17685.

38 C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng and S. P. Ong, A Critical Review of Machine Learning of Energy Materials, *Adv. Energy Mater.*, 2020, **10**, 1903242.

39 A. Mahmood and J. L. Wang, Machine learning for high performance organic solar cells: current scenario and future prospects, *Energy Environ. Sci.*, 2021, **14**, 90–105.

40 X. Rodríguez-Martínez, E. Pascual-San-José and M. Campoy-Quiles, Accelerating organic solar cell material's discovery: high-throughput screening and big data, *Energy Environ. Sci.*, 2021, **14**, 3301–3322.

41 Z. Zhao, Y. Geng, A. Troisi and H. Ma, Performance Prediction and Experimental Optimization Assisted by Machine Learning for Organic Photovoltaics, *Adv. Intell. Syst.*, 2022, **4**, 2100261.

42 T. Lu, M. Li, W. Lu and T. Y. Zhang, Recent progress in the data-driven discovery of novel photovoltaic materials, *J. Mater. Inf.*, 2022, **2**, 7.

43 A. Mahmood, A. Irfan and J.-L. Wang, Machine Learning for Organic Photovoltaic Polymers: A Minireview, *Chin. J. Polym. Sci.*, 2022, **40**, 870–876.

44 X. Rodríguez-Martínez, E. Pascual-San-José, Z. Fei, M. Heeney, R. Guimerà and M. Campoy-Quiles, Predicting the photocurrent–composition dependence in organic solar cells, *Energy Environ. Sci.*, 2021, **14**, 986–994.

45 M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger and C. J. Brabec, Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10% Energy-Conversion Efficiency, *Adv. Mater.*, 2006, **18**, 789–794.

46 Z. Zhao, M. del Cueto, Y. Geng and A. Troisi, Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells, *Chem. Mater.*, 2020, **32**, 7777–7787.

47 B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar and J. M. Buriak, How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics, *ACS Nano*, 2018, **12**, 7434–7444.

48 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 and C. J. Brabec, Elucidating the Full Potential of OPV Materials Utilizing a High-Throughput Robot-Based Platform and Machine Learning, *Joule*, 2021, **5**, 495–506.

49 N. G. An, J. Y. Kim and D. Vak, Machine learning-assisted development of organic photovoltaics via high-throughput in situ formulation, *Energy Environ. Sci.*, 2021, **14**, 3438–3446.

50 H. Sahu, W. Rao, A. Troisi and H. Ma, Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors, *Adv. Energy Mater.*, 2018, **8**, 1801032.

51 Y. Wen, Y. Liu, B. Yan, T. Gaudin, J. Ma and H. Ma, Simultaneous Optimization of Donor/Acceptor Pairs and Device Specifications for Nonfullerene Organic Solar Cells Using a QSPR Model with Morphological Descriptors, *J. Phys. Chem. Lett.*, 2021, **12**, 4980–4986.

52 S. Kar, N. Sizochenko, L. Ahmed, V. S. Batista and J. Leszczynski, Quantitative structure-property relationship model leading to virtual screening of fullerene derivatives: Exploring structural attributes critical for photoconversion efficiency of polymer solar cell acceptors, *Nano Energy*, 2016, **26**, 677–691.

53 E. Abbasi Jannat Abadi, H. Sahu, S. M. Javadpour and M. Goharimanesh, Interpretable machine learning for developing high-performance organic solar cells, Mater, *Today Energy*, 2022, **25**, 100969.

54 J. Munshi, W. Chen, T. Chien and G. Balasubramanian, Transfer Learned Designer Polymers for Organic Solar Cells, *J. Chem. Inf. Model.*, 2021, **61**, 134–142.

55 S. P. Peng and Y. Zhao, Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors, *J. Chem. Inf. Model.*, 2019, **59**, 4993–5001.

56 S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik and C. J. Brabec, Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems, *Adv. Mater.*, 2020, **32**, 1907801.

57 R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt and A. Aspuru-Guzik, Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics, *Energy Environ. Sci.*, 2011, **4**, 4849–4861.

58 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.

59 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 and A. Aspuru-Guzik, Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry-the Harvard Clean Energy Project, *Energy Environ. Sci.*, 2014, **7**, 698–704.

60 E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery, *Adv. Funct. Mater.*, 2015, **25**, 6495–6502.

61 S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares and A. Aspuru-Guzik, Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics, *Joule*, 2017, **1**, 857–870.

62 C. Zanlorenzi and L. Akcelrud, Theoretical studies for forecasting the power conversion efficiencies of polymer-based organic photovoltaic cells, *J. Polym. Sci., Part B: Polym. Phys.*, 2017, **55**, 919–927.

63 Y. Imamura, M. Tashiro, M. Katouda and M. Hada, Automatic High-Throughput Screening Scheme for Organic Photovoltaics: Estimating the Orbital Energies of Polymers from Oligomers and Evaluating the Photovoltaic Characteristics, *J. Phys. Chem. C*, 2017, **121**, 28275–28286.

64 N. Li, I. McCulloch and C. J. Brabec, Analyzing the efficiency, stability and cost potential for fullerene-free organic photovoltaics in one figure of merit, *Energy Environ. Sci.*, 2018, **11**, 1355–1361.

65 E. Abbasi Jannat Abadi, H. Sahu and S. M. Javadpour, M. Goharimanesh, Interpretable machine learning for developing high-performance organic solar cells, *Mater. Today Energy*, 2022, **25**, 100969.

66 S. Nagasawa, E. Al-Naamani and A. Saeki, Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest, *J. Phys. Chem. Lett.*, 2018, **9**, 2639–2646.

67 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 and K.-H. Wei, Enhancing photovoltaic performance by tuning the domain sizes of a small-molecule acceptor by side-chain-engineered polymer donors, *J. Mater. Chem. A*, 2019, **7**, 072–3082.

68 F.-C. Chen, and Virtual Screening of Conjugated Polymers for Organic Photovoltaic Devices Using Support Vector Machines and Ensemble Learning, *Int. J. Polym. Sci.*, 2019, **2019**, 4538514, DOI: 10.1155/2019/4538514.

69 D. Padula, J. D. Simpson and A. Troisi, Combining electronic and structural features in machine learning models to predict organic solar cells properties, *Mater. Horiz.*, 2019, **6**, 343–349.

70 H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang and H. Ma, Designing promising molecules for organic solar cells via machine learning assisted virtual screening, *J. Mater. Chem. A*, 2019, **7**, 17480–17488.

71 H. Sahu and H. Ma, Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning, *J. Phys. Chem. Lett.*, 2019, **10**, 7277–7284.

72 M. Lee, Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design, *Adv. Energy Mater.*, 2019, **9**, 1900891.

73 M.-H. Lee, A Machine Learning–Based Design Rule for Improved Open-Circuit Voltage in Ternary Organic Solar Cells, *Adv. Intell. Syst.*, 2010, **2**, 1900108.

74 Y. Lin, J. Wang, Z. G. Zhang, H. Bai, Y. Li, D. Zhu and X. Zhan, An electron acceptor challenging fullerenes for efficient polymer solar cells, *Adv. Mater.*, 2015, **27**, 1170–1174.

75 W. Zhao, S. Li, H. Yao, S. Zhang, Y. Zhang, B. Yang and J. Hou, Molecular Optimization Enables over 13% Efficiency in Organic Solar Cells, *J. Am. Chem. Soc.*, 2017, **139**, 7148–7151.

76 P. Cheng, G. Li, X. Zhan and Y. Yang, Next-generation organic photovoltaics based on non-fullerene acceptors, *Nat. Photonics*, 2018, **12**, 131–142.

77 J. Wang, P. Xue, Y. Jiang, Y. Huo and X. Zhan, The principles, design and applications of fused-ring electron acceptors, *Nat. Rev. Chem.*, 2022, **6**, 614–634.

78 C. Yan, S. Barlow, Z. Wang, H. Yan, A. K. Y. Jen, S. R. Marder and X. Zhan, Non-fullerene acceptors for organic solar cells, *Nat. Rev. Mater.*, 2018, **3**, 18003.

79 Y. Wu, J. Guo, R. Sun and J. Min, Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells, *npj Comput. Mater.*, 2020, **6**, 120.

80 M. H. Lee, Robust random forest based non-fullerene organic solar cells efficiency prediction, *Org. Electron.*, 2020, **76**, 105465.

81 M. H. Lee, Identifying correlation between the open-circuit voltage and the frontier orbital energies of non-fullerene organic solar cells based on interpretable machine-learning approaches, *Sol. Energy*, 2022, **234**, 360–367.

82 K. Kranthiraja and A. Saeki, Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells, *Adv. Funct. Mater.*, 2021, **31**, 2011168.

83 Y. Miyake and A. Saeki, Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks, *J. Phys. Chem. Lett.*, 2021, **12**, 12391–12401.

84 K. Kranthiraja and A. Saeki, Machine Learning-Assisted Polymer Design for Improving the Performance of Non-Fullerene Organic Solar Cells, *ACS Appl. Mater. Interfaces*, 2022, **14**, 28936–28944.

85 Y. Miyake, K. Kranthiraja, F. Ishiwari and A. Saeki, Improved Predictions of Organic Photovoltaic Performance through Machine Learning Models Empowered by Artificially Generated Failure Data, *Chem. Mater.*, 2022, **34**, 6912–6920.

86 T. Hao, S. Leng, Y. Yang, W. Zhong, M. Zhang, L. Zhu, J. Song, J. Xu, G. Zhou, Y. Zou, Y. Zhang and F. Liu, Capture the high-efficiency non-fullerene ternary organic solar cells formula by machine-learning-assisted energy-level alignment optimization, *Patterns*, 2021, **2**, 100333.

87 P. Malhotra, S. Biswas, F.-C. Chen and G. D. Sharma, Prediction of non-radiative voltage losses in organic solar

cells using machine learning, *Sol. Energy*, 2021, **228**, 175–186.

88 A. Mahmood and J.-L. Wang, A time and resource efficient machine learning assisted design of non-fullerene small molecule acceptors for P3HT-based organic solar cells and green solvent selection, *J. Mater. Chem. A*, 2021, **9**, 15684–15695.

89 A. Mahmood, A. Irfan and J.-L. Wang, Machine learning and molecular dynamics simulation-assisted evolutionary design and discovery pipeline to screen efficient small molecule acceptors for PTB7-Th-based organic solar cells with over 15% efficiency, *J. Mater. Chem. A*, 2022, **10**, 4170–4180.

90 A. Mahmood, A. Irfan and J. Wang, Developing Efficient Small Molecule Acceptors with sp^2-Hybridized Nitrogen at Different Positions by Density Functional Theory Calculations, Molecular Dynamics Simulations and Machine Learning, *Chem. – Eur. J.*, 2022, **28**, e202103712.

91 X. Liu, Y. Shao, T. Lu, D. Chang, M. Li and W. Lu, Accelerating the discovery of high-performance donor/acceptor pairs in photovoltaic materials via machine learning and density functional theory, *Mater. Des.*, 2022, **216**, 110561.

92 S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, The Harvard organic photovoltaic dataset, *Sci. Data*, 2016, **3**, 160086.

93 E. O. Pyzer-Knapp, G. N. Simm and A. Aspuru Guzik, A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials, *Mater. Horiz.*, 2016, **3**, 226–233.

94 A. Paul, A. Furmanchuk, W. Liao, A. Choudhary and A. Agrawal, Property Prediction of Organic Donor Molecules for Photovoltaic Applications Using Extremely Randomized Trees, *Mol. Inf.*, 2019, **38**, 1900038.

95 N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler and S. P. Russo, Machine learning property prediction for organic photovoltaic devices, *npj Comput. Mater.*, 2020, **6**, 166.

96 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 and K. Sun, Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials, *Sci. Adv.*, 2019, **5**, eaay4275.

97 D. Padula and A. Troisi, Concurrent Optimization of Organic Donor–Acceptor Pairs through Machine Learning, *Adv. Energy Mater.*, 2019, **9**, 1902463.

98 T. W. David, H. Anizelli, P. Tyagi, C. Gray, W. Teahan and J. Kettle, Using Large Datasets of Organic Photovoltaic Performance Data to Elucidate Trends in Reliability Between 2009 and 2019, *IEEE J. Photovoltaics*, 2019, **9**, 1768–1773.

99 T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan and J. Kettle, Enhancing the stability of organic photovoltaics through machine learning, *Nano Energy*, 2020, **78**, 105342.

100 T. W. David and J. Kettle, Design for a Sustainability Approach to Organic Solar Cell Design: the Use of Machine Learning to Quantify the Trade-off between Performance, Stability, and Environmental Impact, *J. Phys. Chem. C*, 2022, **126**, 4774–4784.

101 Z.-W. Zhao, M. del Cueto and A. Troisi, Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors, *Digital Discovery*, 2022, **1**, 266–276.

102 M.-H. Lee, Performance and Matching Band Structure Analysis of Tandem Organic Solar Cells Using Machine Learning Approaches, *Energy Technol.*, 2020, **8**, 1900974.

103 W. Sun, Y. Zheng, Q. Zhang, K. Yang, H. Chen, Y. Cho, J. Fu, O. Odunmbaku, A. A. Shah, Z. Xiao, S. Lu, S. Chen, M. Li, B. Qin, C. Yang, T. Frauenheim and K. Sun, Artificial Intelligence Designer for Highly-Efficient Organic Photovoltaic Materials, *J. Phys. Chem. Lett.*, 2021, **12**, 8847–8854.

104 B. L. Greenstein and G. R. Hutchison, Organic Photovoltaic Efficiency Predictor: Data-Driven Models for Non-Fullerene Acceptor Organic Solar Cells, *J. Phys. Chem. Lett.*, 2022, **13**, 4235–4243.

105 W. Sun, M. Li, Y. Li, Z. Wu, Y. Sun, S. Lu, Z. Xiao, B. Zhao and K. Sun, The Use of Deep Learning to Fast Evaluate Organic Photovoltaic Materials, *Adv. Theory Simul.*, 2019, **2**, 1800116.

106 G. J. Moore, O. Bardagot and N. Banerji, Deep Transfer Learning: A Fast and Accurate Tool to Predict the Energy Levels of Donor Molecules for Organic Photovoltaics, *Adv. Theory Simul.*, 2022, **5**, 2100511.

107 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Model.*, 1988, **28**, 31–36.

108 A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori and M. S. Rao, An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation, *Chem. Res. Toxicol.*, 2020, **33**, 20–37.

109 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, *npj Comput. Mater.*, 2019, **5**, 83.

110 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2012, **12**, 2825–2830.

111 X. Rodríguez-Martínez, E. Pascual-San-José and M. Campoy-Quiles, Accelerating organic solar cell material's discovery: high-throughput screening and big data, *Energy Environ. Sci.*, 2021, **14**, 3301–3322.

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

113 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur,

J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, TensorFlow: A system for large-scale machine learning, in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*, 2016, pp. 265–283, **https://github.com/**.

114 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: *Adv. Neural Inf. Process. Syst.*, 2019. **https://arxiv.org/abs/1912.01703**.

115 J. L. Bredas, Mind the gap!, *Mater. Horiz.*, 2014, **1**, 17–19.

116 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, Reoptimization of MDL keys for use in drug discovery, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.

117 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

118 G. Landrum, RDKit, *Open-Source Cheminformatics*, 2016.

119 R. Fisher, *The Design of Experiments (1935), Edinburgh Oliver Boyd*, 1935.

120 N. Majeed, M. Saladina, M. Krompiec, S. Greedy, C. Deibel and R. C. I. MacKenzie, Using Deep Machine Learning to Understand the Physical Performance Bottlenecks in Novel Thin-Film Solar Cells, *Adv. Funct. Mater.*, 2020, **30**, 1907259.

121 B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning:Generative models for matter engineering, *Science*, 2018, **361**, 360–365.

122 H. Sun, F. Chen and Z. Chen, Recent progress on non-fullerene acceptors for organic photovoltaics, *Mater. Today*, 2019, **24**, 94–118.

123 W. Xu and F. Gao, The progress and prospects of non-fullerene acceptors in ternary blend organic solar cells, *Mater. Horiz.*, 2018, **5**, 206–221.

124 A. Kirkey, E. J. Luber, B. Cao, B. C. Olsen and J. M. Buriak, Optimization of the Bulk Heterojunction of All-Small-Molecule Organic Photovoltaics Using Design of Experiment and Machine Learning Approaches, *ACS Appl. Mater. Interfaces*, 2020, **12**, 54596–54607.

125 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, Mordred: a molecular descriptor calculator, *J. Cheminf.*, 2018, **10**, 4.

126 C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann and E. Willighagen, The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 493–500.

127 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, PubChem substance and compound databases, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.

128 I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-De-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko and I. V. Tetko, Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 533–554.

129 J. Dong, D.-S. Cao, H.-Y. Miao, S. Liu, B.-C. Deng, Y.-H. Yun, N.-N. Wang, A.-P. Lu, W.-B. Zeng and A. F. Chen, ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation, *J. Cheminf.*, 2015, **7**, 60.

130 M. R. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. R. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb and B. Wiswedel, KNIME: The konstanz information miner, *4th Int. Ind. Simul. Conf. 2006*, 2006, 11, 58–61, DOI: **10.1145/1656274.1656280**.

131 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, The WEKA data mining software, *ACM SIGKDD Explor. Newsl.*, 2009, **11**, 10–18.

132 A. Mauri, V. Consonni, M. Pavan and R. Todeschini, DRAGON software: An easy approach to molecular descriptor calculations, *MATCH*, 2006, **56**, 237–248.

133 Q. Zhang, Y. J. Zheng, W. Sun, Z. Ou, O. Odunmbaku, M. Li, S. Chen, Y. Zhou, J. Li, B. Qin and K. Sun, High-Efficiency Non-Fullerene Acceptors Developed by Machine Learning and Quantum Chemistry, *Adv. Sci.*, 2022, **9**, 2104742.

134 B. L. Greenstein, D. C. Hiener and G. R. Hutchison, Computational Evolution of High-Performing Unfused Non-Fullerene Acceptors for Organic Solar Cells, *J. Chem. Phys.*, 2022, **156**, 174107.

135 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J. L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.

136 A. Daina, O. Michielin and V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, *Sci. Rep.*, 2017, **7**, 42717.

137 Z. Shi, Y. Bai, X. Chen, R. Zeng and Z. Tan, Tandem structure: a breakthrough in power conversion efficiency for highly efficient polymer solar cells, *Sustainable Energy Fuels*, 2019, **3**, 910–934.

138 S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik and C. J. Brabec, Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems, *Adv. Mater.*, 2020, **32**, 1907801.

139 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, PubChem: Integrated Platform of Small Molecules and Biological Activities, in *Annu. Rep. Comput. Chem.*, Elsevier BV, 2008, pp. 217–241, DOI: **10.1016/S1574-1400(08)00012-1**.

140 M. de Wergifosse and S. Grimme, Nonlinear-response properties in a simplified time-dependent density

functional theory (sTD-DFT) framework: Evaluation of the first hyperpolarizability, *J. Chem. Phys.*, 2018, **149**, 024108.

141 J. Yan, X. Rodríguez-Martínez, D. Pearce, H. Douglas, D. Bili, M. Azzouzi, F. Eisner, A. Virbule, E. Rezasoltani, V. Belova, B. Dörling, S. Few, A. A. Szumska, X. Hou, G. Zhang, H.-L. Yip, M. Campoy-Quiles and J. Nelson, Identifying structure–absorption relationships and predicting absorption strength of non-fullerene acceptors for organic photovoltaics, *Energy Environ. Sci.*, 2022, **15**, 2958–2973.

142 C. Yan, J. Qin, Y. Wang, G. Li and P. Cheng, Emerging Strategies toward Mechanically Robust Organic Photovoltaics: Focus on Active Layer, *Adv. Energy Mater.*, 2022, **12**, 2201087.