ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Review

Machine Learning in Non-fullerene Organic Solar Cells: Accelerating Discovery, Design, and Understanding

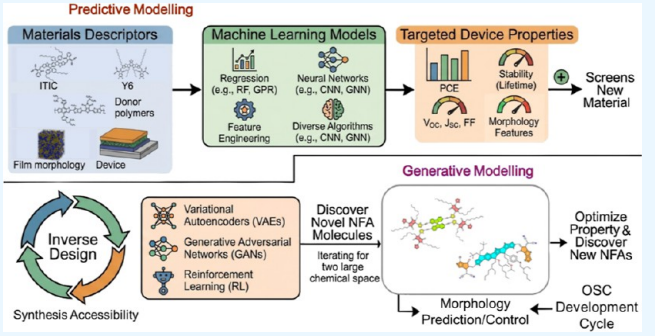
Bibhas Das and Anirban Mondal*

Cite This: https://doi.org/10.1021/acsomega.6c01194

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The emergence of nonfullerene acceptors (NFAs) has transformed the performance of organic solar cells (OSCs), driving laboratory power conversion efficiencies beyond 19%. Yet the immense combinatorial space of possible donor−acceptor materials renders exhaustive experimental exploration impractical. Machine learning (ML) and artificial intelligence have therefore become powerful tools for accelerating materials discovery, performance prediction, and molecular design in organic photovoltaics. This review provides a critical and systematic assessment of predictive and generative ML approaches applied to OSCs, encompassing methods ranging from random forests and gradient boosting to deep learning architectures such as graph neural networks, transformers, variational autoencoders, generative adversarial networks, and genetic algorithms. We examine advances in molecular representation and feature engineering, evaluate model performance in predicting key device metrics—including power conversion efficiency, open-circuit voltage, short-circuit current density, and fill factor—and assess the extent to which computational predictions translate into experimentally validated devices. We further discuss key limitations, including data set bias, distribution shift, morphology-related effects, chemical validity, and the synthetic accessibility of generated candidates, as well as the persistent gap between predicted and realized performance. Finally, we highlight emerging directions poised to shape the next phase of ML-guided OSC research, including physics-informed learning, multiobjective optimization, interpretable models, active learning coupled with first-principles calculations, and human-in-the-loop discovery pipelines. Collectively, these advances position machine learning as a central component in the rational design and accelerated development of next-generation organic photovoltaic materials.



1. INTRODUCTION

Organic solar cells (OSCs) have emerged as a promising class of photovoltaic technologies due to their intrinsic advantages, including mechanical flexibility, low weight, solution processability at low temperatures, and the ability to tailor optical and electronic properties through molecular design. These attributes position OSCs as strong candidates for applications such as building-integrated photovoltaics, flexible electronics, and wearable energy systems, where conventional inorganic photovoltaics face intrinsic limitations.[1−5] In addition, compatibility with roll-to-roll manufacturing and simplified device architectures offers a viable pathway toward cost-effective and scalable energy production.[6−14] Against the backdrop of escalating global energy demand and sustainability imperatives, OSCs continue to attract substantial research attention as a next-generation renewable energy platform.

A defining milestone in the evolution of OSCs occurred in 2015 with the advent of nonfullerene acceptors (NFAs), marked by the introduction of ITIC by Han and co-workers.[15,16] This development initiated a rapid departure from fullerene-based acceptors and triggered an unprecedented rise in power conversion efficiencies (PCEs), with state-of-the-art NFA-based devices now exceeding 20% under laboratory conditions.[17−23] Such progress has narrowed the performance gap between OSCs and established inorganic photovoltaic technologies, underscoring the transformative impact of molecular acceptor design.

The success of ITIC and subsequent NFAs is rooted in their acceptor−donor−acceptor (A−D−A) molecular architecture, typically comprising a fused aromatic core such as indacenodithiophene (IDT) and strong electron-withdrawing terminal groups like 1,1-dicyanomethylene-3-indanone (IC). This design paradigm enables (i) systematic tuning of frontier molecular orbital energies to optimize open-circuit voltage (V_{OC}) while

Received: January 31, 2026
Revised: April 24, 2026
Accepted: May 5, 2026

minimizing interfacial energetic losses, (ii) strong and broad-band absorption extending into the near-infrared region, and (iii) rigid, planar backbones that promote favorable molecular packing and morphological stability.[24−26] Subsequent molecular innovations rapidly built upon this framework. For example, fluorinated derivatives such as IT-4F reduced voltage losses through enhanced intermolecular interactions. At the same time, the Y6 family introduced highly electron-deficient fused cores that simultaneously optimized absorption, charge transport, and nanoscale morphology, yielding record efficiencies beyond 19%.[27,28]

Despite these advances, critical challenges remain that impede the commercial deployment of high-efficiency OSCs. Bulk heterojunction morphologies are governed by complex, multi-scale structure−property relationships that are difficult to predict from molecular structure alone. Persistent voltage losses arising from nonradiative recombination, susceptibility to photo-oxidative degradation, and long-term morphological instability continue to limit operational stability and device lifetimes.[29−34] Addressing these challenges requires not only new molecular designs but also fundamentally improved strategies for navigating the immense chemical design space of OSC materials.

The chemical design space of NFAs is extraordinarily vast. Contemporary NFAs span more than 50 distinct core scaffolds, encompassing both fused and nonfused architectures, a wide range of electron-withdrawing terminal groups, and an effectively unbounded diversity of side-chain substitutions.[35−41] When combined with the already extensive library of polymer donors, the number of possible donor−acceptor pairs reaches into the billions. Traditional experimental discovery strategies—based on iterative synthesis, device fabrication, and characterization—are therefore inherently inefficient and increasingly untenable, relying heavily on trial-and-error and chemical intuition.[36−38]

Machine learning (ML) and artificial intelligence (AI) offer a decisive shift away from this Edisonian paradigm toward data-driven discovery. By learning quantitative structure−property and structure−performance relationships from existing data sets, ML models can predict key photovoltaic metrics directly from molecular representations, prioritize promising candidates prior to synthesis, and guide experimental efforts with substantially higher success rates.[36,37,39−56] Importantly, these approaches do not merely accelerate screening; they reshape the conceptual framework of materials discovery by enabling prospective design, hypothesis-driven exploration of chemical space, and iterative feedback between computation and experiment. More recently, the role of ML in OSC research has expanded beyond performance prediction toward inverse molecular design. Generative models capable of proposing entirely new NFA structures optimized for multiple objectives—such as efficiency, stability, and synthetic accessibility—are emerging as powerful tools for next-generation materials discovery. This evolution marks a critical transition from predictive modeling to autonomous or semiautonomous molecular design.

In this review, we provide a comprehensive and critical assessment of machine learning and artificial intelligence applications in organic solar cells, with a particular focus on nonfullerene acceptors. We first discuss the data landscape underpinning ML-driven OSC research, including available data sets, data quality, representation strategies, and inherent limitations. We then examine predictive ML models for photovoltaic performance, spanning classical regression techniques to modern deep learning architectures, and evaluate their practical impact and shortcomings. Subsequently, we review generative ML approaches—such as variational autoencoders, generative adversarial networks, transformer-based models, and genetic algorithms—that enable inverse molecular design and chemical space exploration.

This review is intended as a critical and representative synthesis rather than a formal meta-analysis. To clarify the scope of coverage, we prioritize studies that focus on nonfullerene-acceptor-based OSCs or directly inform their machine-learning workflows, and that explicitly report data set construction, molecular representation, model architecture, benchmarking, interpretability, generative design, or experimental validation. The literature discussed here has been identified through structured searches using combinations of keywords such as “organic solar cells”, “non-fullerene acceptors”, “machine learning”, “deep learning”, “graph neural network”, “transformer”, “generative model”, and “inverse design”, while foundational OPV and molecular-ML studies are retained where necessary to establish core concepts or contextualize data resources.

Throughout the review, we identify key challenges that remain unresolved, including ensuring chemical validity and synthetic feasibility of generated molecules, capturing morphology-dependent effects, and bridging the gap between computational predictions and experimental validation. Finally, we outline future directions for the field, emphasizing physics-informed machine learning, multiobjective optimization, improved model interpretability, and human-in-the-loop workflows that integrate ML with domain expertise to enable reliable and accelerated discovery of next-generation OSC materials.

2. DATA AND MOLECULAR REPRESENTATIONS FOR MACHINE LEARNING IN ORGANIC SOLAR CELLS

2.1. Data Resources: Computational and Experimental Data Sets

The development of reliable machine learning models for organic solar cells is fundamentally constrained by the availability, quality, and diversity of underlying data. Because device performance emerges from complex and multiscale structure−property relationships, effective data sets must encode both molecular-level electronic features and experimentally measured photovoltaic metrics. Existing data resources can be broadly categorized into computational and experimental data sets, which play complementary roles in model training, validation, and generalization.

Among computational resources, the Harvard Clean Energy Project Database (CEPDB) remains the most comprehensive data set available for organic photovoltaic materials.[57,58] CEPDB and CEPDB-derived data sets contain quantum-chemically derived electronic properties for very large libraries of candidate OPV molecules. The original CEPDB release reported approximately 2.3 million molecular structures, whereas more recent NFA-focused studies often rely on smaller, curated subsets of around 51,000 computational NFAs. These data sets typically include HOMO and LUMO energies, optical gaps, and related quantum-chemical descriptors computed using density functional theory (DFT), predominantly at the B3LYP/6-31G(d) level or higher, or via closely related computational workflows.[57−59] Owing to its scale and chemical diversity, CEPDB has become a cornerstone for data-driven OSC research, particularly for pretraining deep learning models. In

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

Table 1. Summary of Representative Datasets Frequently Used in Machine-Learning Studies of Organic Solar Cells[a]

data set	approx. size	data source/content	typical use case	known limitations
CEPDB/CEPDB-derived computational libraries[57−59]	2.3 million structures (full CEPDB); 51,256 NFAs (DeepAcceptor subset)	large-scale quantum-chemical OPV data set; DFT-derived electronic descriptors and model-estimated photovoltaic labels	pretraining, transfer learning, descriptor learning, and large-scale virtual screening	computational (not experimental) ground truth; simplified PCE estimation; lacks morphology, processing, and interfacial realism; potential donor/chemistry bias in older subsets leading to distribution shift vs experiments
Wu et al. benchmark data set[64]	565 experimentally reported D/A combinations	literature-curated donor−acceptor devices with reported PCE; widely used for benchmarking early ML models	benchmarking classical ML models (RF, BRT, SVR, ANN) and enabling prospective validation	modest size; early generation chemical space; limited standardized processing metadata; weaker coverage of out-of-distribution chemistries
Suthar et al. experimental device data set[56]	1242 experimentally verified donor/NFA combinations	literature-curated polymer/NFA devices with PCE, V_{OC}, J_{SC}, FF, frontier-orbital descriptors, and cheminformatic features	supervised prediction of device metrics, descriptor comparison, and SHAP analysis	still small for deep learning; heterogeneous fabrication and measurement conditions across studies; morphology and processing only indirectly represented
DeepAcceptor experimental data set[59]	1027 small-molecule NFAs from 508 articles, paired with a 51,256-molecule computational pretraining set	literature-curated acceptor structures with experimental PCE labels plus a curated computational NFA library	fine-tuning abcBERT and acceptor-focused screening/generation	donor identity ignored during aggregation; repeated molecules reduced to the maximum reported PCE; acceptor-centric labels compress device and processing variability into a single value
Liu et al. PUFp database[62]	1343 experimental NFA-related OPV entries with 260 donor materials	literature-derived OPV acceptor database represented with polymer-unit fingerprints and auxiliary descriptors	structure−property analysis, fingerprint benchmarking, and SHAP-guided design of acceptors	material- and pair-level reporting are partially aggregated; metadata are curated from literature rather than standardized remeasurement; processing realism remains uneven across entries
OSC-Net multifidelity data set[65]	47,329 computational entries +1782 experimental data points	combined low-fidelity computational data and high-fidelity literature/lab experimental measurements; repeated experimental entries retained	multifidelity learning, uncertainty quantification, and prediction of V_{OC}, J_{SC}, FF, and PCE	computational subset uses simplified physics and limited donor diversity; strong fidelity gap between computational and experimental domains; many processing variables remain implicit
Das et al. and Khatua et al. QM-curated experimental data sets[60,61]	400 donor−acceptor devices across two studies	literature-curated OSC donor−acceptor devices with descriptor-rich DFT/TD-DFT annotations, including frontier orbitals, ionization energy/electron affinity, multipolar (dipole/quadrupole), excited-state, exciton-binding, energy-loss, and interfacial charge-transfer descriptors	supervised prediction of V_{OC}, J_{SC}, FF, and PCE_{max}; interpretable feature selection; SHAP analysis; SISSO-based equation discovery; physics-informed ML	moderate size; literature-derived (not uniformly remeasured); heterogeneous fabrication and measurement conditions; does not explicitly capture morphology, processing history, or full device-stack effects

[a]Sizes are reported as provided in the original sources or in the associated curated subsets.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

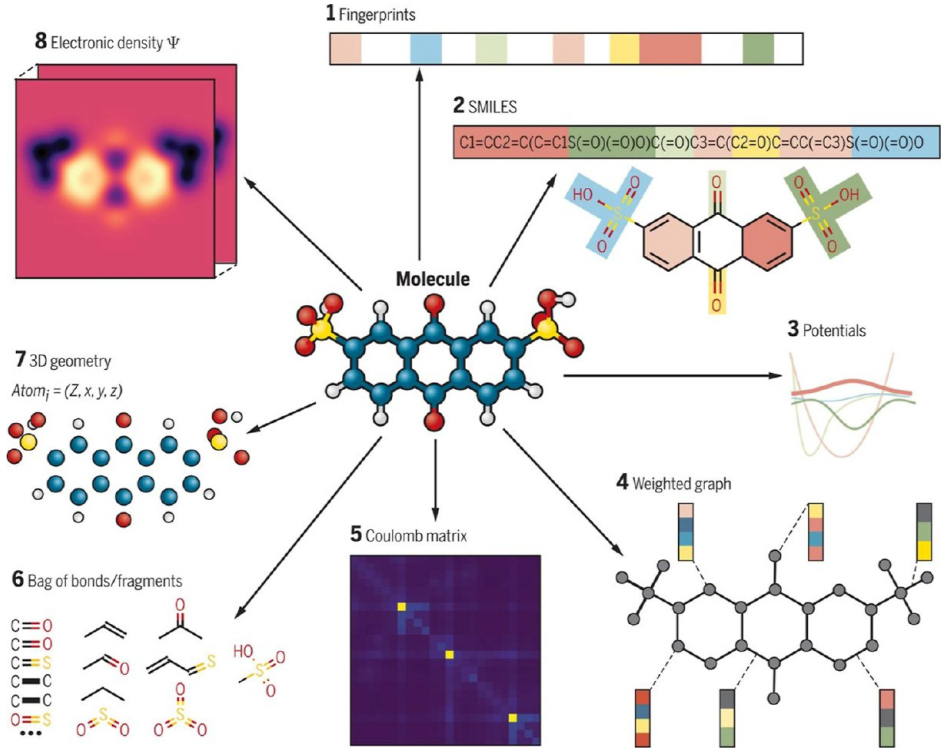


Figure 1. Common molecular representation paradigms derived from a single molecule (1) molecular fingerprints; (2) SMILES string encoding; (3) learned or physics-based potential energy features; (4) weighted molecular graphs; (5) Coulomb matrix representation; (6) bag-of-bonds or fragment-based descriptors; (7) three-dimensional atomic geometry; (8) electronic density distributions. Reprinted with permission from ref 66. Copyright 2025 American Association for the Advancement of Science.

this context, transfer learning strategies leverage CEPDB to enable neural networks to learn generic structure−property correlations, which can subsequently be fine-tuned using smaller experimental data sets containing ground-truth device performance metrics.[58]

Despite its utility, CEPDB exhibits inherent limitations that must be carefully considered. The molecular structures are computationally generated rather than experimentally synthesized, and reported PCE values are derived from simplified, model-based estimations rather than direct device measurements. Consequently, CEPDB does not capture critical effects arising from morphology, processing conditions, or interfacial physics, and models trained exclusively on this data set may exhibit limited predictive fidelity when applied to real devices.

In addition to such widely used public data sets, our own recent studies by Das and Mondal[60] and Khatua et al.[61] contribute a combined quantum-mechanically curated data set of 400 experimentally reported donor−acceptor devices across two complementary studies, comprising a 300-device data set for supervised prediction of J_{SC}, V_{OC}, FF, and PCE_{max}, and a 100-device data set designed for physics-informed modeling and data-driven equation discovery. These data sets are distinctive in that they move beyond routine cheminformatic descriptors and incorporate a chemically interpretable descriptor space derived from DFT/TD-DFT calculations, including donor- and acceptor-resolved frontier orbital energies (HOMO, HOMO − 1, LUMO, and LUMO + 1), dipole moments, acceptor quadrupole moments, orbital-energy splitting terms, donor ionization energy, acceptor electron affinity, optical gaps/absorption descriptors, oscillator strengths, exciton-binding energies, energy loss (E_{loss}), ground-to-excited-state dipole-moment differences ($\Delta\mu$), and composite interfacial energetics such as ΔE^{D-A}_{HOMO}, ΔE^{D-A}_{LUMO}, ΔG_{CR}, ΔG_{CS}, and ΔG_{CT}. The resulting descriptor sets are particularly valuable for interpretable learning, feature attribution, and physics-guided model construction, as they directly connect molecular electronic structure to device-relevant observables. At the same time, these data sets remain moderate in size, are literature-curated rather than generated under a unified experimental protocol, and do not explicitly encode morphology, processing history, or full device-stack effects; they should therefore be viewed as high-fidelity, descriptor-rich resources for mechanistic modeling rather than exhaustive benchmarks for universal generalization.

In contrast, experimental data sets provide direct access to measured device performance and therefore serve as indispensable benchmarks for evaluating ML model accuracy and practical relevance. Although substantially smaller, several curated experimental data sets have recently emerged as key resources. For example, Liu et al. compiled a data set of 1343 experimentally characterized donor−acceptor pairs and reported comprehensive photovoltaic metrics, including PCE, open-circuit voltage, short-circuit current density (J_{SC}), and fill factor (FF).[62] Sun et al. reported a data set of 1027 small-molecule NFA structures paired with experimentally measured PCE values, specifically designed to support deep learning workflows.[59] Similarly, the data set assembled by Suthar et al. comprises 1242 donor−acceptor combinations spanning diverse chemical families and emphasizes systematic evaluation of molecular descriptor choices for supervised learning.[56] While these data sets are orders of magnitude smaller than computational repositories, they provide irreplaceable ground truth for validating ML predictions against experimentally realized device behavior.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

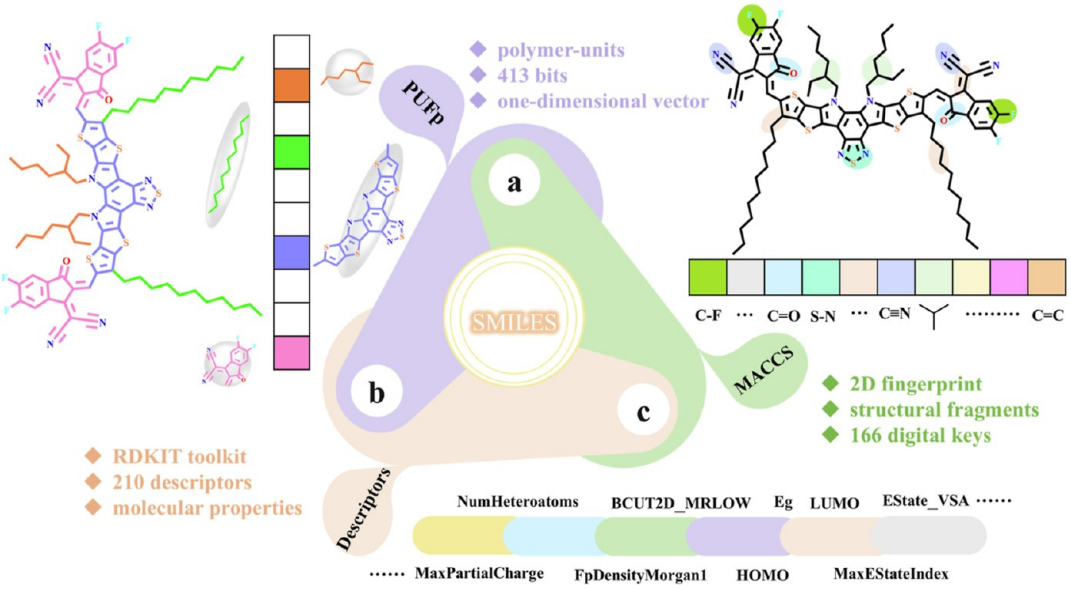


Figure 2. Molecular representation strategies are commonly used in machine learning models for organic electronic materials. (a) Polymer unit fingerprinting (PUFp), where each bit corresponds to the presence of a specific polymer unit encoded via one-hot representation; (b) RDKit-derived molecular descriptors capturing physicochemical properties; (c) MACCS fingerprints, representing predefined molecular substructures using binary encoding. Reprinted with permission from ref 62. Copyright 2025 Springer Nature (npj Computational Materials).

A persistent challenge in assembling experimental data sets lies in ensuring data consistency and reliability. OSC performance metrics are reported across laboratories that employ different device architectures, fabrication protocols, and measurement conditions, leading to substantial variability and potential systematic bias. As a result, rigorous data curation and cleansing procedures have become increasingly important. Recent studies have implemented standardized filtering protocols to remove entries with incomplete metadata, implausible values, or inconsistent reporting, thereby improving data set integrity and downstream model generalization.[63] Although such preprocessing steps are labor-intensive, they are essential for constructing robust ML models capable of delivering reproducible and physically meaningful predictions.

For clarity, the major data-related limitations discussed throughout this review are consolidated here. Current OSC machine learning data sets exhibit a pronounced fidelity gradient. Large computational libraries, such as CEPDB and CEPDB-derived subsets, are valuable for pretraining and broad chemical-space exploration; however, they rely on computed descriptors and model-derived device labels that do not capture morphology, processing history, or interfacial losses under realistic device conditions. In contrast, experimental data sets provide the most relevant ground truth for key performance metrics (PCE_{max}, V_{OC}, J_{SC}, and FF), yet they remain comparatively small and heterogeneous. This heterogeneity arises because nominally similar donor−acceptor systems are often reported under varying solvents, additives, annealing protocols, layer thicknesses, device architectures, and measurement conventions across different laboratories.

This combination of limited experimental data, literature-selection bias toward high-performing systems, and the resulting shift in distribution between the computational and experimental domains constitutes a central challenge in the field. As a consequence, models that achieve strong retrospective performance on curated data sets may still exhibit limited reliability in prospective or experimentally realizable settings. Table 1 summarizes representative data sets widely used in OSC machine-learning studies, along with their typical roles and recurring limitations.[56,59,62−65]

2.2. Feature Engineering: Molecular Descriptors and Learned Representations

Effective machine learning for organic solar cells critically depends on feature representations that encode the molecular and mesoscale characteristics governing photovoltaic performance. Because OSC efficiency arises from an interplay of electronic structure, molecular packing, and processing-dependent morphology, no single descriptor class is universally sufficient. Consequently, a diverse spectrum of feature engineering strategies has been developed, ranging from physically motivated electronic descriptors to data-driven learned representations (see Figure 1).

2.2.1. Frontier Molecular Orbital Descriptors. Electronic structure descriptors such as HOMO and LUMO energy levels, optical band gaps (E_g), and reorganization energies constitute one of the earliest and most widely adopted feature sets in OSC-focused ML models. These quantities are typically computed using density functional theory, most commonly at the B3LYP/6-31G(d) level of theory.[36,40,41,56] Frontier orbital descriptors directly relate to key photovoltaic parameters, including open-circuit voltage and charge transport, and have demonstrated strong predictive utility across numerous studies. However, their scope is inherently limited to the electronic properties of isolated molecules. It does not account for intermolecular interactions, morphology, or processing-induced structural effects that play decisive roles in device performance.

2.2.2. Fingerprint-Based Molecular Representations. Molecular fingerprints encode chemical structures into fixed-length vectors that conventional ML algorithms can readily process. MACCS keys comprise 167 binary features indicating the presence or absence of predefined substructures, providing a compact, computationally efficient representation. Extended Connectivity Fingerprints (ECFPs) provide a more flexible framework by iteratively encoding atom-centered neighbor-

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

hoods over increasing radii, thereby capturing local chemical environments at multiple length scales. More recently, Polymer-Unit Fingerprints (PUFs) have been introduced to better reflect the chemistry of conjugated materials by fragmenting molecules into chemically meaningful building blocks and encoding fragment presence using one-hot schemes (see Figure 2). Zhu et al. demonstrated that PUFs outperform traditional MACCS keys for conjugated organic semiconductors, highlighting the importance of domain-specific fingerprint design for OSC applications.[62]

2.2.3. Fragment-Based Molecular Fingerprints. Beyond predefined fingerprints, fragment-based representations have emerged as a powerful means of encoding chemically interpretable information. Zhang et al. developed a representation strategy that systematically fragments donor and nonfullerene acceptor molecules according to minimal-ring and molecular-unit principles, then encodes each fragment as a binary feature.[67] The resulting high-dimensional fingerprints reflect the complete fragment vocabulary observed across the data set. Machine learning models trained on these representations—such as random forest and extra trees regressors—exhibited strong predictive performance for photovoltaic metrics. Importantly, SHapley Additive exPlanations (SHAP) analysis enabled identification of fragments with the greatest impact on predicted efficiency. By recombining high-importance fragments above a similarity threshold, this framework was further extended to generate millions of virtual donor and acceptor candidates for large-scale screening.

2.2.4. Coulomb Matrices and Atomic Descriptors. Coulomb matrices and related atomic descriptors encode three-dimensional molecular geometry by representing pairwise electrostatic interactions between atoms as a function of nuclear charges and interatomic distances.[68] These representations preserve spatial information that is absent from purely connectivity-based fingerprints and are particularly valuable in ML models targeting quantum-chemical properties. While sensitive to molecular conformation and atom indexing, various normalization and sorting schemes have been proposed to mitigate these limitations, enabling their application in geometry-aware neural network architectures.

2.2.5. Sub-Unit Electronic Descriptors. An intermediate approach between whole-molecule descriptors and atomistic representations involves decomposing molecules into chemically meaningful subunits and assigning electronic descriptors to each fragment. Zhang et al. demonstrated that partitioning nonfullerene acceptors into a small number of aromatic subunits and representing each molecule using the electronic properties of these components significantly improves the prediction of energy levels.[63] This strategy reduces the reliance on full-molecule DFT calculations while retaining chemically interpretable features. Notably, the study also emphasized rigorous data cleansing to eliminate inconsistent experimental records prior to training, underscoring the strong coupling between descriptor quality and data integrity.

2.2.6. Morphological Descriptors. To bridge the gap between molecular structure and device-level performance, several studies have incorporated morphological descriptors that capture the complexity of the bulk heterojunction active layer. These include $\pi-\pi$ stacking distances, donor−acceptor interfacial separations, blend compositions, solvent and additive choices, and thermal annealing conditions.[69−72] Inclusion of such features typically improves predictive accuracy by 5−15%, highlighting their critical role in determining charge generation and transport. However, these descriptors often require molecular dynamics (MD) simulations or extensive experimental characterization, limiting their scalability and widespread applicability.

2.2.7. Graph Neural Network Representations. Graph neural networks (GNNs) represent a shift toward end-to-end learned representations that eliminate the need for manual feature engineering. In these models, molecules are represented as graphs with atoms as nodes and bonds as edges. Architectures such as SchNet incorporate continuous-filter convolutions based on interatomic distances while preserving physical invariances, whereas message-passing neural networks (MPNNs) and graph convolutional networks (GCNs) iteratively aggregate local structural information to construct hierarchical feature embeddings.[73−77] By learning task-specific representations directly from molecular graphs, GNN-based models frequently outperform handcrafted descriptors, particularly when trained on sufficiently large and diverse data sets.

2.3. Interpretability and Feature Importance in Machine Learning Models

Beyond predictive accuracy, the practical utility of machine learning models in organic solar cell research depends critically on their interpretability. Identifying which molecular features most strongly influence photovoltaic performance is essential for validating model behavior against established chemical intuition and, more importantly, for translating ML outputs into actionable design principles.

Early efforts to assess feature importance in OSC-focused ML models primarily relied on tree-based algorithms such as random forest regression, which provide intrinsic measures of feature relevance through impurity reduction or permutation importance. These studies consistently revealed that descriptors associated with fused-ring cores, aromatic heterocycles, and conjugation length dominate power conversion efficiency predictions.[78] Such trends are chemically intuitive, as the molecular scaffold dictates frontier orbital energies, absorption profiles, and intermolecular packing tendencies, all of which play central roles in governing charge generation and transport within bulk heterojunction devices.[60,78,79] More recently, model-agnostic interpretability techniques—most notably Shapley Additive Explanations—have enabled a more rigorous and quantitative assessment of descriptor contributions. SHAP values decompose individual predictions into additive contributions from each feature by averaging over all possible feature coalitions, thereby providing a theoretically grounded and consistent framework for feature attribution.[56,60,61] When applied to OSC performance-prediction models, SHAP analyses repeatedly identify frontier molecular orbital energies (HOMO and LUMO), optical band gaps, and descriptors of molecular polarity, size, and symmetry as the most influential variables governing PCE. These results reinforce the central role of electronic structure while also highlighting the importance of molecular shape and dipolar characteristics in determining charge separation and recombination dynamics.[80−84]

A notable implication of these interpretability studies is that a relatively small subset of physically meaningful descriptors often captures a significant fraction of the variance in device performance. This observation suggests that aggressive dimensionality reduction and informed feature selection may simplify ML models without compromising predictive accuracy, improving robustness and generalization—particularly for experimentally derived data sets that are limited in size.[60]

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

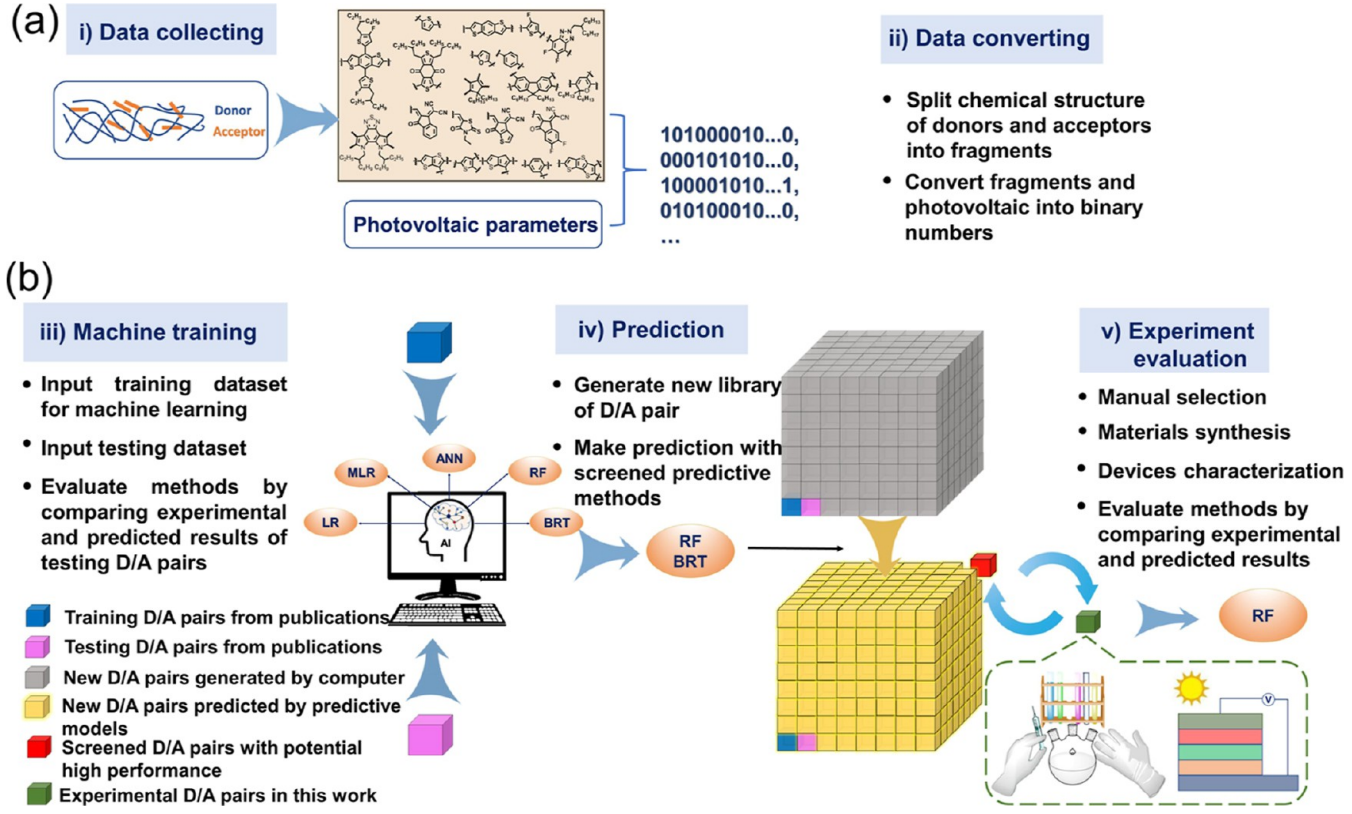


Figure 3. (a) Scheme of collecting experimental data and converting chemical structures into digitized input suitable for machine learning. (b) Scheme of training machine learning models, predicting target properties, and evaluating model performance via experimentation. Reprinted with permission from ref 64. Copyright 2025 Springer Nature (npj Computational Materials).

From a materials design perspective, interpretable feature importance analyses provide a valuable bridge between data-driven prediction and rational molecular engineering, enabling researchers to prioritize structural motifs and electronic characteristics that are most likely to yield high-performance organic photovoltaic materials.[56,60]

3. PREDICTIVE MACHINE LEARNING FOR ORGANIC SOLAR CELL PERFORMANCE

3.1. Tree-Based Machine Learning Models for OSC Performance Prediction

Early efforts to predict organic solar cell performance relied on simple, physically motivated empirical models. A notable example is the Scharber model, which correlates power conversion efficiency with linear combinations of the donor HOMO energy, the acceptor LUMO energy, and the optical bandgap. While conceptually appealing, this approach yields weak correlations with experimental data. It systematically underestimates the complexity of OSC operation by neglecting critical factors such as bulk heterojunction morphology, nonradiative recombination pathways, and interfacial electronic structure effects.[85,86]

The introduction of classical machine learning methods—particularly ensemble tree-based models—marked a significant advance over such empirical formulations. Random Forest (RF) regression and gradient boosting regression (GBR) models demonstrated a substantially improved capacity to capture the nonlinear and multivariate relationships governing OSC performance. In a representative study, Wu et al. trained RF models on a data set of 565 donor−acceptor pairs and achieved strong predictive accuracy, with R^2 values of 0.84 on independent test sets. This performance exceeded that of linear regression, support vector regression, and conventional artificial neural networks evaluated under comparable conditions.[64] Importantly, RF-based feature importance analysis revealed that descriptors associated with molecular scaffold and core structure dominate PCE predictions, reinforcing experimental insights into the central role of acceptor and donor backbone design.[64,87]

The practical relevance of these classical ML approaches was further demonstrated through prospective experimental validation. In the same study, six donor−acceptor combinations predicted by the RF model to exhibit high efficiency were synthesized and fabricated into devices, yielding measured PCE values within 1−2 percentage points of the model predictions (see Figure 3). Such close agreement between prediction and experiment provided early and compelling evidence that ML models could be used not only for retrospective analysis but also to guide the selection of experimental materials.[64]

Gradient boosting methods have since emerged as particularly effective for predicting OSC performance. Across multiple studies, gradient boosting models routinely achieve coefficients of determination exceeding $R^2 \approx 0.85$ and mean absolute errors of 1.5−2.5 PCE percentage points on held-out test sets. Their superior performance relative to random forests is commonly attributed to the sequential learning strategy inherent to boosting, in which successive decision trees are trained to correct the residual errors of preceding models. This iterative refinement enables gradient boosting to more effectively model highly nonlinear structure−performance relationships, espe-

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

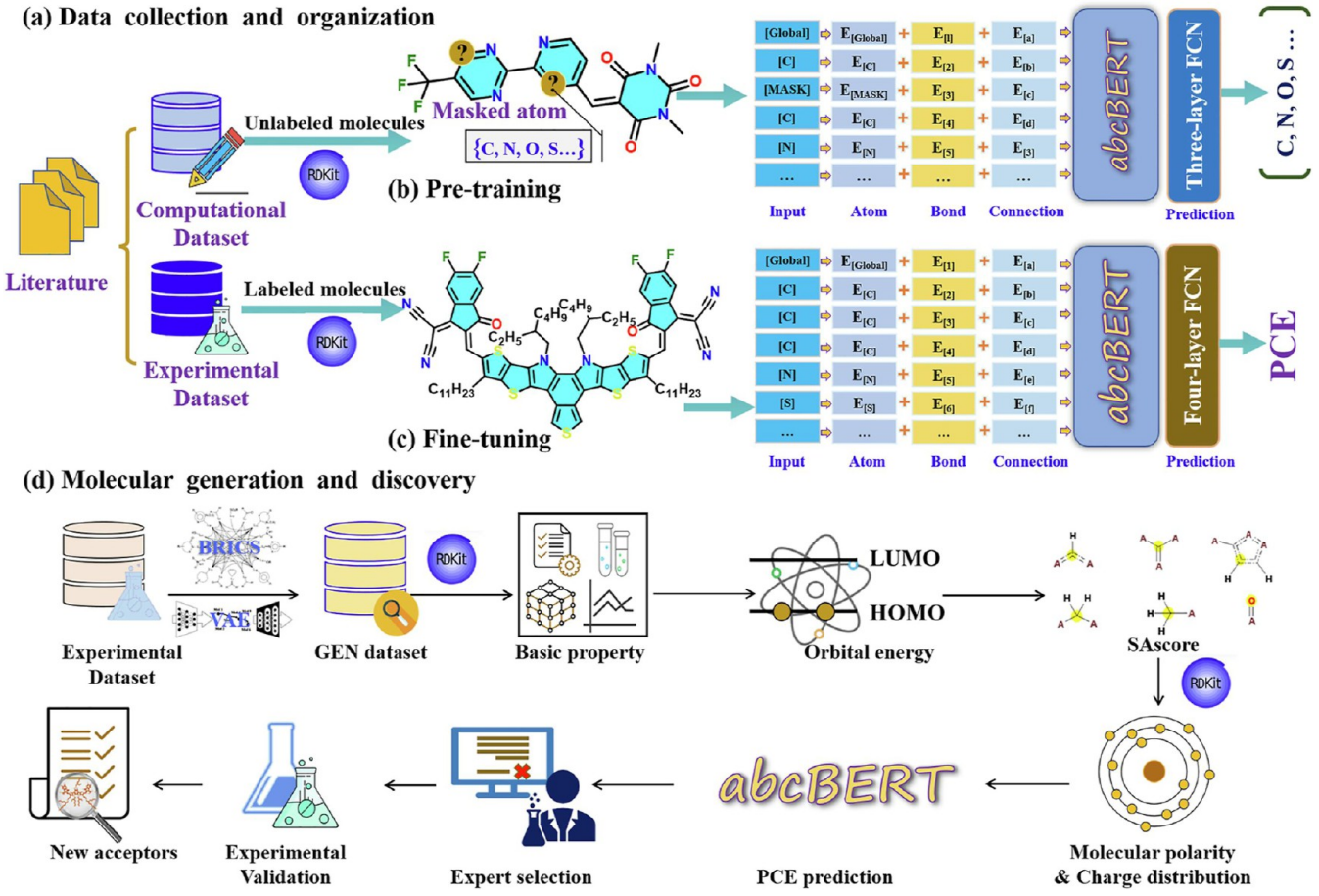


Figure 4. Overview of the DeepAcceptor framework (a) computational and experimental acceptor data collected from the literature served as unlabeled and labeled molecules; (b) the abcBERT model was pretrained by predicting masked atoms of unlabeled molecules with the assistance of bond lengths and connections; (c) the pretrained model was fine-tuned on the experimental data set; (d) a molecular generation and screening process was built to identify high-performance acceptor candidates. Reprinted with permission from ref 59. Copyright 2025 Springer Nature (npj Computational Materials).

cially in heterogeneous data sets typical of experimentally reported OSC measurements.[60,88]

3.2. Deep Learning Architectures for Performance Prediction and Molecular Discovery

As data set sizes have grown and molecular representations have become more expressive, deep learning architectures have increasingly surpassed classical machine learning methods in predicting organic solar cell performance. Unlike descriptor-driven models, deep learning approaches enable end-to-end learning directly from molecular graphs, strings, or fragment representations, allowing the model to autonomously extract task-relevant features without manual engineering.

Transformer-based architectures have emerged as particularly powerful in this context. A prominent example is the abcBERT model developed by Sun et al., which encodes atom types, bond connectivity, and spatial relationships using transformer layers.[59] The model was first pretrained on 51,256 molecules from the Harvard Clean Energy Project database and subsequently fine-tuned on a curated experimental data set of 1027 NFA−PCE pairs. This transfer learning strategy yielded a mean absolute error of 1.8% and R^2 of 0.94 on held-out test data, representing a substantial improvement over classical ensemble methods. Beyond retrospective prediction, abcBERT demonstrated prospective discovery capability by screening millions of candidate acceptors and identifying multiple molecules predicted to exceed 14% PCE. Several of these candidates were synthesized and experimentally validated, with the highest-performing molecule achieving a PCE of 14.61%, providing compelling evidence that transformer-based models can guide real-world materials discovery (see Figure 4).

Generative deep learning models further extend predictive frameworks by enabling inverse molecular design. Variational autoencoders (VAEs) are particularly attractive because they can encode molecular structures into continuous latent spaces from which novel molecules can be sampled. Sun et al. introduced DeepAcceptor, a semisupervised VAE framework based on SELFIES fragments and augmented with graph-BERT predictors, enabling the generation of chemically valid donor−acceptor pairs with experimentally verified PCEs approaching 14%. By coupling latent-space exploration with predictive filtering, such frameworks demonstrate how deep generative models can transition from passive prediction to active materials design.[59]

Convolutional neural networks (CNNs) offer an alternative deep learning paradigm, particularly for fingerprint-, image-, or string-based molecular representations. In these models, molecular structures—encoded as 2D diagrams, molecular fingerprint bitmaps, or SMILES strings—are treated as grid-like inputs, enabling convolutional layers to extract localized patterns associated with chemical substructures and photo-

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

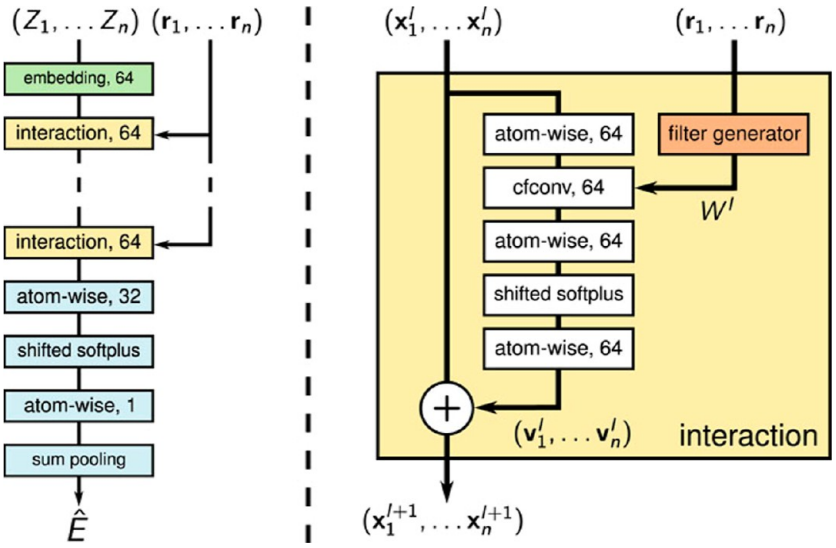


Figure 5. Illustrations of the SchNet architecture and its interaction blocks with the atom embedding in green, interaction blocks in yellow, and the property prediction network in blue, alongside the detailed structure of the continuous-filter convolution filter-generating network in orange. Reprinted with permission from ref 89. Copyright 2025 American Institute of Physics.

voltaic performance. Early studies showed that CNNs could successfully predict NFA properties and generate candidate molecules, often augmented with attention mechanisms to improve interpretability and fragment attribution, and validated using quantum chemical calculations. Chen et al. employed CNN-based architectures to design 12,224 novel donor−acceptor pairs predicted to exceed 19% PCE, reporting a maximum predicted efficiency of 19.20%.[76] However, experimental validation of such high-efficiency predictions remains challenging, as several candidates either failed to achieve predicted performance or proved synthetically inaccessible.

Hybrid deep learning architectures combining residual networks and attention mechanisms have also shown promise. The SolarPCE-Net architecture developed by Liu et al. integrates residual connections to stabilize training of deep networks with self-attention layers that selectively weight different molecular regions and capture coupling effects between donor and acceptor components.[62] These models demonstrate improved generalization relative to traditional descriptor-based approaches, particularly when extrapolating to chemical spaces far removed from the training distribution. The combination of deep residual learning and attention-based feature weighting provides a robust framework for modeling the highly nonlinear and interdependent structure−performance relationships characteristic of organic photovoltaic materials.

3.3. Graph Neural Networks and Message-Passing Architectures

Graph neural networks (GNNs) have emerged as a powerful paradigm for molecular property prediction in organic solar cells, primarily due to their ability to operate directly on molecular graph representations. In these models, atoms are treated as nodes and chemical bonds as edges, allowing the network to learn hierarchical representations that encode both local chemical environments and global molecular topology. By avoiding reliance on predefined molecular descriptors, GNNs can capture subtle structure−property relationships that are often inaccessible to fingerprint-based approaches. By avoiding reliance on predefined descriptors, GNNs can internalize features such as conjugation patterns and noncovalent interactions that strongly influence OSC performance.[74]

Among early physics-informed architectures, SchNet has attracted significant attention within the computational chemistry community. SchNet employs continuous-filter convolutional layers that operate on interatomic distances, while explicitly enforcing translational and rotational invariance. This symmetry-aware design ensures that learned representations respect fundamental physical constraints, improving generalizability across chemically diverse molecular systems (see Figure 5). Such properties are particularly advantageous for OSC materials, where subtle variations in conjugation length, functionalization, and noncovalent interactions can lead to pronounced differences in photovoltaic performance.[73,89]

In another study, Wang et al.[90] develop an efficient screening framework that integrates a graph neural network (SLI-GNN) with an ensemble LightGBM model to directly predict OSC power conversion efficiencies from molecular structure. The SLI-GNN is pretrained on approximately 200,000 CEPDB entries to predict a small set of key microscopic molecular properties from graph inputs, which are then used as inputs to the LightGBM "efficiency" model trained on high-quality experimental PCE data. This hybrid architecture establishes quantitative links between structural features, molecular properties, and device performance while eliminating the need for costly DFT calculations during screening, enabling rapid high-throughput evaluation of candidate molecules. The framework was validated against experimental data and highlights the importance of careful feature selection—identifying nine critical properties—to bridge microscopic descriptors and macroscopic OSC efficiency, demonstrating how deep learning combined with ensemble learning can accurately link chemical structure to device performance and significantly accelerate OSC materials discovery.

Message-passing neural networks (MPNNs) provide a more general framework in which atomic representations are iteratively updated by aggregating information from neighboring atoms through learned message functions. With each message-passing step, the receptive field of an atom expands, enabling the model to capture progressively larger chemical environments.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

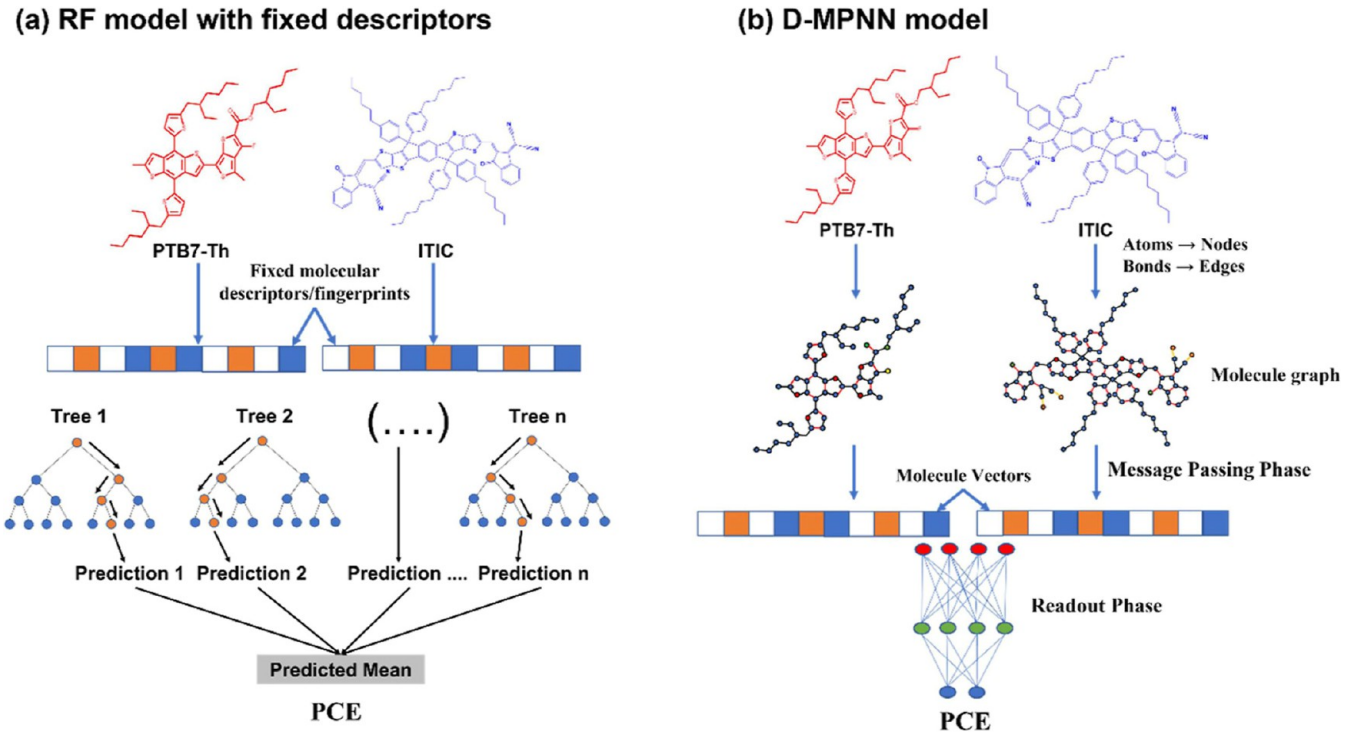


Figure 6. Comparison of modeling workflows: (a) conventional Random Forest models employing fixed molecular descriptors (RDKit, Mordred) and fingerprints (Morgan); (b) directed message-passing neural networks implemented using the `Chemprop` framework. Reprinted with permission from ref 75. Copyright 2025 American Association for the Advancement of Science.

This iterative aggregation naturally reflects the multiscale origin of molecular properties, which arise from an interplay between local bonding motifs and extended molecular architecture.[75] Applications of GNNs to OSC performance prediction have consistently demonstrated superior accuracy compared to traditional descriptor-based models. Zhang et al.[91] developed a GNN-based high-throughput screening framework trained on a large data set of molecular structures with associated photovoltaic properties. Their model achieved a mean absolute error of approximately 0.3% for PCE prediction and an R^2 exceeding 0.9, enabling the identification of several previously unexplored candidate materials with predicted PCE values above 15%. Subsequent experimental validation confirmed the predictive reliability of the model, highlighting the practical utility of graph-based learning for materials discovery.

More recent studies have further refined message-passing strategies for OSC applications. Directed message-passing neural networks (D-MPNNs), in which messages propagate along directed chemical bonds, have proven particularly effective. By explicitly encoding bond directionality, D-MPNNs mitigate information redundancy and improve representation efficiency (see Figure 6). Malhotra et al. demonstrated that D-MPNNs trained directly on donor−acceptor pairs significantly outperform models based on fixed molecular descriptors and fingerprints, underscoring the advantages of end-to-end learned graph representations for predicting OSC device performance.[75]

3.4. Transfer Learning and Pretraining Strategies

Despite their strong representational capacity, deep learning models for OSC prediction are often constrained by the limited size of experimentally curated data sets. Transfer learning has therefore emerged as a critical strategy for improving model robustness and generalization.[92,93] In this approach, models are first pretrained on large computational data sets and subsequently fine-tuned on smaller experimental data sets relevant to OSC device performance.[94]

The Clean Energy Project Database, containing over 51,000 computationally generated organic molecules with associated electronic properties, has become a cornerstone resource for such pretraining efforts. Neural networks trained on CEPDB learn transferable representations of molecular structure, conjugation patterns, and electronic features without requiring costly experimental data. These pretrained models can then be fine-tuned on experimental data sets comprising only 1000−2000 donor−acceptor systems, yielding substantially improved performance compared to models trained from scratch.[59,94]

The success of this strategy reflects the observation that many structure−property relationships governing OSC behavior—such as trends in frontier orbital energies, molecular size, and conjugation—are largely transferable across chemical contexts. Pretraining enables models to internalize general principles, while fine-tuning adapts them to the more complex, noisy experimental domain. In practice, transfer learning has been shown to improve predictive accuracy by approximately 5−10% on experimental test sets, representing a meaningful advance in the reliability and applicability of deep learning models for OSC materials design.[59,94] A recent multifidelity study (OSC-Net) employs a two-step pipeline in which models are pretrained on 47,329 computational entries and subsequently fine-tuned on 1782 experimental measurements. The results demonstrate that sequentially combining low-fidelity computational data with high-fidelity experimental data yields superior predictive performance compared to single-fidelity models. OSC-Net predicts V_{OC}, J_{SC} and FF from 4096-bit Morgan fingerprints of donor and acceptor plus the donor/acceptor ratio, computes PCE from those outputs for screening, and adds uncertainty

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

quantification and replicate experimental entries to capture batch variability and strengthen practical relevance. The model was validated against published experiments and used for high-throughput candidate selection. Importantly, the authors report a clear domain shift (computational median PCE ≈ 1.06% vs experimental median PCE ≈ 4.88%), which motivates fine-tuning of pretrained models to correct biases in computational estimates.[65]

3.5. Predicting Component Performance Metrics: V_{OC}, J_{SC}, and FF

While overall power conversion efficiency remains the most commonly predicted metric, decomposing device performance into its constituent components—open-circuit voltage, short-circuit current density, and fill factor—provides deeper insight into the underlying device physics and enables more targeted optimization strategies, distinct physical processes govern each of these metrics: V_{OC} is primarily influenced by donor−acceptor energy level alignment and nonradiative recombination losses; J_{SC} reflects light absorption efficiency, exciton dissociation, and charge collection; and FF is sensitive to charge transport balance and recombination dynamics under operating conditions.[55,95−98]

Machine learning models trained to predict these individual parameters generally exhibit lower predictive accuracy than those targeting PCE directly, with typical R^2 values in the range of 0.75−0.85 compared to 0.85−0.95 for PCE.[60,81,99,100] This reduction reflects the greater sensitivity of component metrics to processing conditions, morphology, and interfacial effects that are often not explicitly encoded in molecular descriptors. Nevertheless, accurate component-level predictions enable more sophisticated design strategies. For example, materials predicted to exhibit high V_{OC} but limited J_{SC} can be systematically modified to enhance optical absorption while preserving favorable voltage characteristics. The ability to predict V_{OC}, J_{SC}, and FF simultaneously thus represents a significant methodological advance, shifting ML models from purely empirical performance predictors toward tools that can support mechanism-aware optimization and rational molecular design in organic solar cells.

3.6. Strengths and Limitations of Predictive Approaches

3.6.1. Strengths. Predictive machine learning models offer several compelling advantages for accelerating research and development in organic solar cells. Most notably, they enable rapid, low-cost screening of vast chemical spaces without requiring experimental synthesis or device fabrication, thereby compressing discovery timelines by an order of magnitude or more. By providing quantitative performance estimates, these models guide experimental prioritization and allow finite laboratory resources to be focused on the most promising candidate materials.

Beyond acceleration, ML models facilitate the extraction of structure−property relationships that may remain obscured in conventional trial-and-error workflows. Feature attribution methods, such as SHAP analysis and attention-weight visualization, have increasingly enabled the identification of chemically meaningful motifs governing photovoltaic performance. Notably, several studies have demonstrated successful prospective validation, in which ML-predicted high-efficiency materials were synthesized and experimentally confirmed to achieve power conversion efficiencies close to the ML-predicted values. Such demonstrations provide strong evidence that predictive models can move beyond retrospective fitting toward genuinely predictive and discovery-oriented tools.[56,59,64]

3.6.2. Limitations. Despite these strengths, several fundamental limitations continue to constrain the practical impact of current predictive approaches. Foremost among these is the issue of limited extrapolation capability. Most models perform optimally within the chemical space spanned by their training data, but their prediction accuracy deteriorates rapidly under distribution shift. This limitation is particularly problematic for materials discovery, where the most transformative candidates are often structurally dissimilar from previously studied systems.[101−105]

A second major limitation is the inability of structure-only models to capture morphological effects. Device-scale performance in OSCs is critically governed by nanoscale phase separation, domain purity, and percolation pathways—features that are emergent properties of molecular ensembles rather than isolated molecules. Accurately incorporating morphology typically requires molecular dynamics simulations or multiscale modeling frameworks, which remain computationally expensive and challenging to integrate seamlessly into high-throughput ML pipelines.[60,106−112]

Interpretability presents an additional challenge. While post hoc analysis tools provide useful correlations, most predictive models remain fundamentally black-box and do not directly encode physical constraints into the learning process. As a result, models may generate predictions that violate basic physical principles when extrapolating beyond their training domain, such as open-circuit voltages exceeding donor−acceptor bandgap limits or short-circuit currents surpassing theoretical absorption maxima. These failures highlight the need for physics-informed or constraint-aware learning frameworks.[36,41,84,113,114]

Finally, computational cost remains a nontrivial bottleneck in many workflows. Descriptor-based models that rely on quantum chemical calculations often require tens of minutes to hours per molecule at the DFT level, significantly limiting the throughput of large-scale screening efforts. While deep learning models operating directly on molecular graphs partially alleviate this constraint, developing reliable, physically grounded, and computationally efficient descriptor-free approaches remains an open challenge.

Collectively, these strengths and limitations underscore both the transformative potential and the current boundaries of predictive machine learning in OSC research. Addressing these challenges will require tighter integration of physical principles, multiscale modeling, and uncertainty-aware learning strategies to enable robust extrapolation and truly predictive materials discovery.

4. GENERATIVE MACHINE LEARNING FOR INVERSE MOLECULAR DESIGN

4.1. Rule-Based and Evolutionary Strategies for Molecular Generation

The earliest generative strategies for organic solar cell materials relied on rule-based molecular construction and genetic algorithms (GAs), which explicitly encode chemical intuition and evolutionary search principles to explore candidate design spaces. In these approaches, molecules are assembled from predefined fragments or motifs, and successive generations are optimized using fitness functions tied to target photovoltaic metrics (see Figure 7).

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

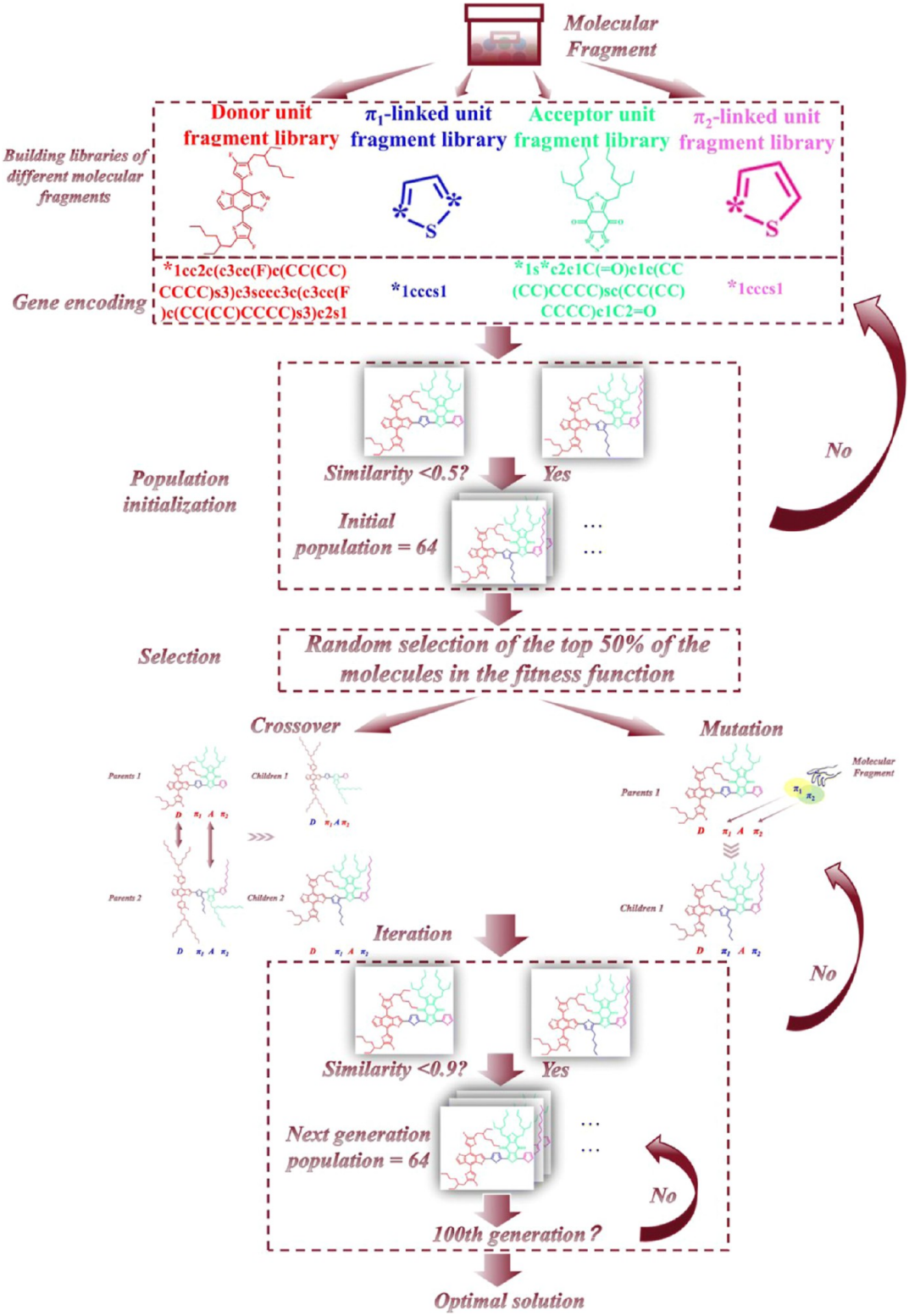


Figure 7. Schematic illustration of the genetic algorithm (GA) optimization process, including population initialization, fitness evaluation, selection, crossover, mutation, and iteration until termination. Reprinted with permission from ref 115. Copyright 2025 American Association for the Advancement of Science.

A seminal example is the work of Greenstein and Hutchison, who systematically screened more than 5000 unfused nonfullerene acceptor structures generated via fragment recombination. By evaluating these candidates against predicted optoelectronic properties, they identified robust design rules linking core-end-group arrangements to enhanced power conversion efficiencies exceeding 18%. Extending this framework, iterative GA-driven optimization for tandem device architectures revealed evolutionary pathways favoring broad absorption and complementary energy-level alignment, offering

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

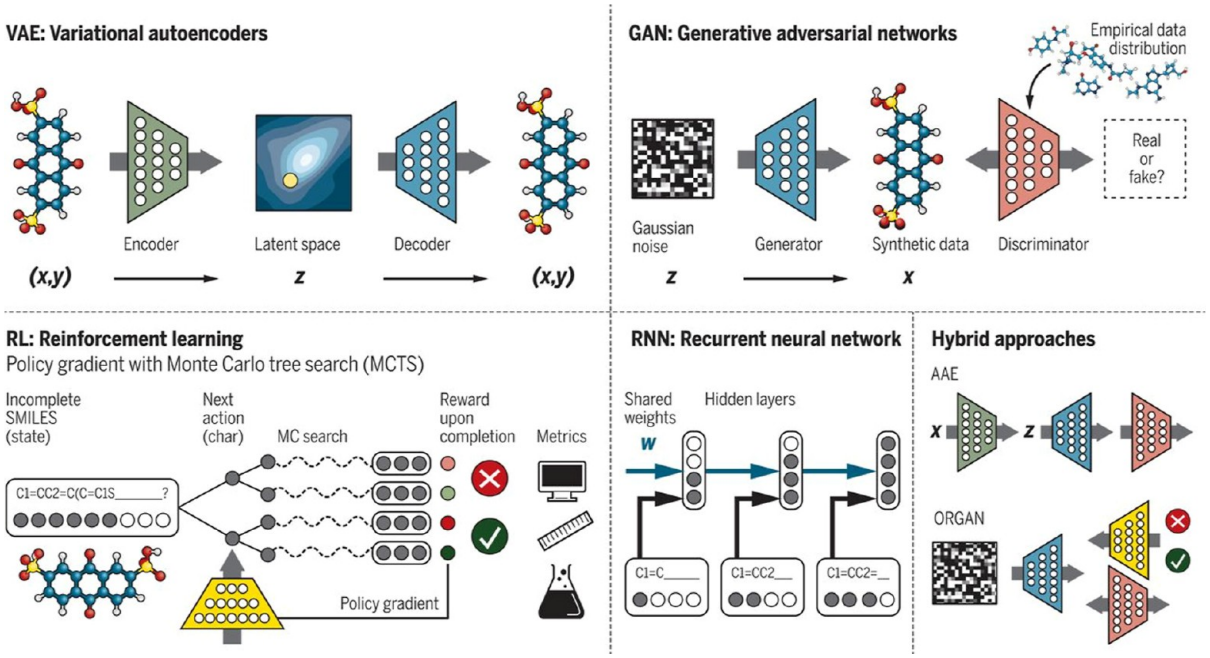


Figure 8. Schematic representation of several generative model architectures commonly used for inverse design and molecule/material generation, including variational autoencoders (VAEs), generative adversarial networks (GANs), reinforcement learning (RL), recurrent neural networks (RNNs) and hybrid models. Reprinted with permission from ref 66. Copyright 2025 American Association for the Advancement of Science.

mechanistic insight into how high-performing absorbers emerge within constrained chemical spaces. Although these early generative approaches did not employ data-driven representation learning, they established several enduring principles. Most importantly, they demonstrated that large libraries of chemically valid OSC candidates can be constructed and navigated efficiently using modular building blocks and heuristic optimization strategies. Moreover, their transparent rule-based nature provided interpretable structure−property relationships that continue to inform fragment selection, fitness function design, and chemical constraints in modern generative frameworks. As such, rule-based and GA-driven methods laid the conceptual groundwork for contemporary deep generative models in inverse molecular design.[105]

4.2. Deep Generative Architectures for Molecular Design

Advances in deep learning have enabled a new class of generative approaches capable of learning continuous, high-dimensional molecular representations and sampling novel structures beyond predefined fragment libraries (see Figure 8). Variational autoencoders (VAEs) are particularly well suited for molecular generation, as they encode chemical structures into continuous latent spaces from which new molecules can be stochastically sampled and decoded. This probabilistic formulation facilitates smooth interpolation between known molecules and provides implicit uncertainty quantification through reconstruction likelihoods, offering a measure of confidence in generated candidates.[66,116,117]

In OSC discovery workflows, VAEs are frequently combined with chemically informed fragmentation schemes to enhance validity and diversity.[59] For example, latent-space sampling coupled with BRICS-based recombination has enabled the generation of millions of chemically plausible donor and acceptor candidates. Rigorous evaluation of validity, uniqueness, and novelty metrics ensures that the generated molecular libraries maintain chemical realism prior to downstream property prediction and screening.[59,63,67,78,115,118,119]

Generative adversarial networks (GANs) provide an alternative generative paradigm based on adversarial training between a generator and a discriminator. When properly trained, GANs can generate molecular structures that closely resemble the distributions of real compounds, yielding high structural diversity and chemical plausibility. However, GANs often suffer from training instability and mode collapse, which can limit their practical deployment relative to VAEs in molecular design tasks.[66,117,120] Conditional generative models further extend these frameworks by incorporating explicit property constraints—such as target HOMO/LUMO energies, optical band gaps, or polarity descriptors—directly into the generation process. By conditioning molecular generation on desired photovoltaic attributes, these models enable targeted exploration of chemically relevant regions of design space, substantially reducing the need for post hoc filtering and iterative optimization. Such property-aware generation represents a critical step toward truly inverse molecular design for organic solar cell materials.[77,117]

4.3. Transformer-Based Molecular Generation

Transformer architectures, initially developed for natural language processing, have been successfully adapted for molecular generation by treating chemical structures as sequences, most commonly using SMILES or related string-based encodings. In this paradigm, molecules are generated token by token using self-attention mechanisms that capture long-range dependencies and global structural context. Autoregressive transformer models such as GPT-2 can be fine-tuned on molecular corpora to learn the underlying grammar of chemical sequences and subsequently generate novel SMILES strings that are decoded into molecular structures.[121−124]

A key advantage of transformer-based generators lies in their expressive attention mechanisms, which enable modeling of nonlocal correlations in molecular topology that are difficult to capture with purely local graph-based approaches. In addition, transformers readily accommodate conditional generation by

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

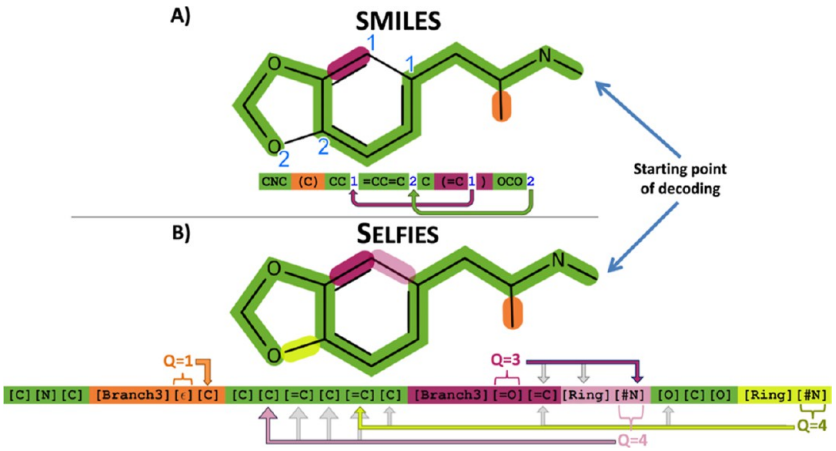


Figure 9. Illustration of representing a molecular graph (here 3,4-methylenedioxymethamphetamine) with two computer-friendly, string-based molecular encoding schemes: (A) derivation of the molecular graph using SMILES, where the main string (green) is augmented with branches (enclosed in brackets) and rings (annotated by matching ring closure digits); (B) derivation using SELFIES, where each token is defined such that atomic valence constraints are inherently satisfied and all symbolic information (except ring closure) is local under the encoding rules. Reproduced from Krenn et al., ref 126, licensed under CC BY 4.0. Published by IOP Publishing Ltd.

appending property tokens or auxiliary embeddings, enabling biasing toward target electronic properties relevant to OSC performance. Their compatibility with large-scale pretraining further enhances data efficiency, particularly when experimental data sets are limited.[121−124]

Despite these strengths, sequence-based generative models suffer from a fundamental limitation: syntactic correctness does not guarantee chemical validity. A significant fraction of generated SMILES strings correspond to invalid or chemically implausible structures, necessitating extensive postgeneration filtering and validation. While alternative encodings such as SELFIES partially mitigate this issue, ensuring high validity remains a central challenge for transformer-driven molecular generation.[125,126]

4.4. Molecular String Encodings: SMILES and SELFIES

The molecular string representation used in generative models plays a decisive role in determining both the validity and efficiency of molecular generation. The Simplified Molecular Input Line Entry System (SMILES) encodes chemical structures as linear strings using atom symbols, bond types, and ring closures. Although SMILES is widely adopted and readily interpretable, it lacks inherent syntactic and chemical constraints. As a result, many SMILES strings generated by sequence-based models are invalid or chemically implausible, necessitating extensive postgeneration validation and filtering that can significantly reduce effective yield.[125,126]

To address these limitations, Krenn et al.[126] introduced SELFIES (Self-Referencing Embedded Strings), a robust molecular encoding scheme that guarantees 100% chemical validity by construction. SELFIES employs a constrained grammar in which every possible string corresponds to a chemically valid molecule, thereby eliminating the generation of invalid structures (see Figure 9). This property has made SELFIES particularly attractive for deep generative models, especially transformer- and VAE-based architectures, as it enables direct exploration of chemical space without costly post hoc correction steps. Consequently, SELFIES-based encodings have become increasingly prevalent in generative workflows for OSC materials discovery, substantially improving the efficiency and reliability of inverse molecular design pipelines.[66,125]

4.5. End-to-End Generative Pipelines for OSC Discovery

Recent studies increasingly emphasize end-to-end generative pipelines that integrate large-scale virtual chemical space enumeration with data-driven screening, physics-informed filtering, and experimental feasibility constraints. Fragment-based strategies have proven remarkably scalable. Zhang et al.[67] combined fragment-level molecular fingerprints with Random Forest and Extra Trees regression to model photovoltaic performance across curated donor-nonfullerene acceptor data sets. Systematic fragment recombination enabled the virtual construction of over 24 billion donor−acceptor combinations, from which high-throughput screening identified candidates with predicted PCEs approaching 13.2%, demonstrating that fragment-level ML pipelines can effectively navigate enormous chemical spaces while retaining predictive reliability. Explainability-driven workflows further extend this paradigm. In a subsequent study, Zhang et al.[78] curated 547 experimentally validated donor−acceptor pairs. They employed Morgan and MACCS fingerprints with Random Forest regression and used SHAP to identify efficiency-determining substructures. Guided recombination of high-importance donors, acceptors, and π-bridges yielded a virtual library of approximately 3.45 billion donor−acceptor pairs. From this space, over 14,000 candidates exhibited predicted PCEs above 14%, and 123 exceeded 15.5%, with a maximum predicted efficiency of ∼15.9%. These results illustrate how interpretability can be leveraged not only for post hoc analysis but as an active design driver.

Complementary deep-learning-driven pipelines developed by Cao, Lv, and co-workers[118] reinforce these findings. Cao et al. implemented two parallel strategies: an LSTM-based PCE predictor coupled with fragment recombination, generating approximately 7600 novel donor−acceptor pairs with predicted PCEs above 18%, and a genetic-algorithm framework combined with Random Forest regression using 43 structural descriptors, yielding candidates with predicted efficiencies up to 16.85%.[115] In related work, Lv et al.[82] integrated generative LSTM models with symbolic regression to balance predictive accuracy and interpretability. Their workflow produced over 200,000 donors and 870,000 acceptors, which were combinatorially assembled into a library of ∼185 billion pairs. Screening identified 5753 candidates with predicted PCEs above 18.5%, while symbolic

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

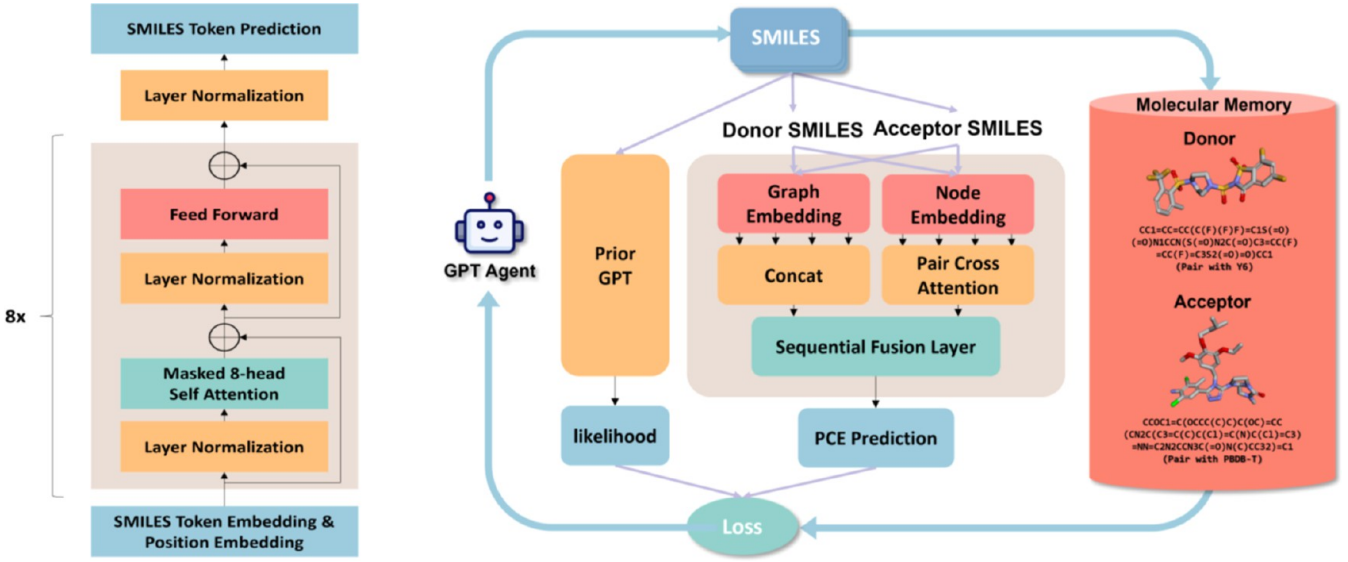


Figure 10. Illustration of the generative pretraining and reinforcement learning (RL) model architectures used for data-driven molecular design: (a) the GPT-based generative pretrained transformer architecture with stacked self-attention layers and autoregressive token prediction; (b) the structure of the RL agent, including policy and value networks interacting with an environment to optimize molecular design objectives. Figure adapted from Qiu et al., ref 124, arXiv:2503.23766 licensed under a Creative Commons Attribution CC BY 4.0 license.

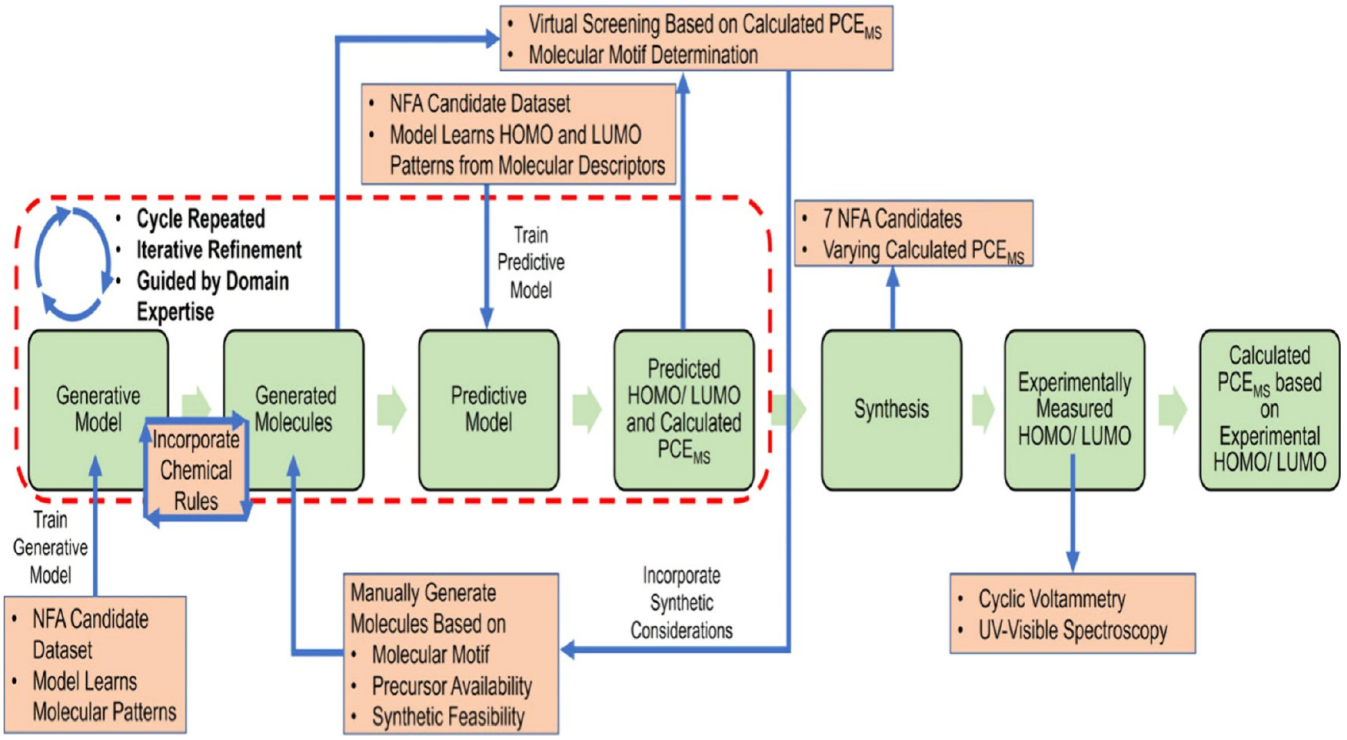


Figure 11. Schematic of the NFA molecular target generation and synthesis workflow, integrating generative machine learning with virtual screening and experimental validation: training a generative model on a library of ∼50,000 NFA candidates to learn molecular patterns and propose novel structures; iterative refinement of generated structures with domain expertise; encoding of molecules as SMILES and prediction of HOMO/LUMO energy levels using a predictive ML model; calculation of PCE_{MS} as a virtual screening criterion to select high-performance NFA candidates; repeated generation and screening cycles to identify molecular motifs; manual design of synthesizable molecules considering precursor availability and synthetic feasibility; synthesis of seven candidate NFAs; and validation of predicted energy levels via CV and UV−vis measurements. Reprinted with permission from ref 127. Copyright 2025 American Chemical Society.

regression uncovered interpretable structure−property relationships inaccessible to purely black-box models. Collectively, these studies suggest that diverse generative strategies can yield comparable outcomes, provided that evaluation and screening are systematic.

Reinforcement learning has further expanded the design space. Qiu et al.[124] employed a transformer-based generative reinforcement learning framework that iteratively biases acceptor generation toward high-PCE regimes using a reward signal derived from predicted efficiency (see Figure 10). This approach

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

identified candidates with predicted PCEs exceeding 21%, highlighting the potential of reward-driven generation to target extreme-performance regions of chemical space. More recently, hybrid pipelines integrating generative enumeration with physics-guided, uncertainty-aware machine learning have gained traction. Das et al.[79] introduced a unified framework combining physics-informed molecular fragmentation, hierarchical clustering, and combinatorial assembly with an evidential message-passing neural network capable of predicting optoelectronic properties—including oscillator strength, LUMO offsets, absorption maxima, and exciton binding energy—while explicitly quantifying uncertainty. Starting from 257 experimentally reported NFAs, this workflow generated a synthetically realistic library of approximately 500,000 acceptor−donor−acceptor (ADA) structures and achieved deterministic enrichment of candidates with tightly converged, target-optimized optoelectronic profiles, validated against quantum chemical calculations.

Finally, Tan et al.[127] demonstrated the importance of human-in-the-loop design by integrating generative modeling, virtual screening, and expert assessment of synthetic feasibility for diketopyrrolopyrrole (DPP)-based NFAs. Selected candidates were synthesized and incorporated into working organic photovoltaic devices, confirming that generative proposals can translate into experimentally realizable materials (see Figure 11). Such workflows underscore a critical shift in the field: successful generative design increasingly depends not only on algorithmic sophistication, but also on integration with physical constraints, uncertainty quantification, and experimental judgment.

4.6. Synthetic Feasibility and Chemical Validity

A persistent bottleneck in generative molecular design is ensuring that algorithmically proposed structures are both chemically valid and synthetically accessible. Molecules optimized purely in silico may suffer from impractical reaction pathways, reliance on exotic intermediates or protecting groups, or intrinsic chemical instability, rendering them unsuitable for experimental realization. Early generative studies frequently encountered this disconnect, producing structurally novel yet synthetically intractable candidates and resulting in low experimental success rates.

Contemporary workflows address this challenge through a combination of algorithmic constraints and expert intervention. Fragment-based generative strategies, particularly those employing BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) decomposition, assemble molecules from chemically meaningful fragments with established synthetic precedents. By recombining known building blocks rather than generating arbitrary structures, these approaches substantially improve the likelihood that proposed molecules can be synthesized using existing methodologies.[63,67,78,115,118,119]

Cheminformatic filtering provides an additional layer of feasibility assessment. Metrics such as molecular weight, lipophilicity, heteroatom count, and hydrogen-bonding capacity—while originally developed in medicinal chemistry—often correlate with synthetic tractability and chemical robustness, enabling rapid exclusion of implausible candidates.[79] Nevertheless, automated filters alone remain insufficient to capture the nuanced judgment required for synthetic feasibility fully. Consequently, close integration of expert synthetic chemists into generative pipelines has emerged as a best practice, enabling human evaluation of reaction plausibility, stability, and scalability that remains difficult to encode computationally.[127]

Chemical and photochemical stability pose related challenges. While quantum chemical calculations can, in principle, evaluate oxidation susceptibility, bond dissociation energies, and excited-state reactivity, such calculations are prohibitively expensive when applied to millions of generated molecules. To address this limitation, recent studies have introduced machine learning surrogate models trained to predict synthetic accessibility scores and stability-related properties. These models enable rapid prescreening of large generative libraries, allowing computationally intensive calculations and experimental validation to be reserved for a small subset of high-confidence candidates.[59]

4.7. Trade-Off-Aware Generative Design

Early generative studies in OSC materials design primarily focused on optimizing a single target—typically power conversion efficiency—without accounting for additional constraints that govern real device viability. In practice, however, high-efficiency materials must also exhibit long-term photochemical stability, synthetic tractability, cost-effectiveness, and environmentally benign processing. These requirements are often mutually competing: structural modifications that enhance absorption or charge transport may simultaneously compromise stability or synthetic simplicity, while morphology-optimizing substituents can adversely affect processability.

Multiobjective optimization frameworks address the challenge of conflicting design criteria by enabling the simultaneous treatment of several targets, rather than optimizing a single scalar objective.[128] Evolutionary algorithms—particularly genetic algorithms and related population-based schemes—are naturally suited to this setting because they operate on populations of candidates and can accommodate vector-valued or composite fitness functions. Instead of converging to a single optimum, such algorithms can be designed to evolve populations toward sets of solutions that represent trade-offs among device-relevant properties, for example different performance metrics or structural characteristics of organic solar cell morphologies and materials. In practice, many generative and evolutionary pipelines in materials design still employ a scalarized objective (e.g., a weighted combination of predicted properties or microstructure-derived performance metrics),[129,130] but the same formal machinery extends directly to true multiobjective formulations based on Pareto dominance. In a Pareto-based analysis, candidate solutions are compared according to dominance relationships across all objectives, and the nondominated set (the Pareto front) consists of solutions for which no other candidate is simultaneously better in every objective dimension.[128,131] By sampling and analyzing this front, researchers can rationally select materials or microstructures tailored to specific application priorities—for example, choosing morphologies that trade a small loss in short-circuit current for improved robustness or fabrication-related constraints, depending on the design context.[131]

The formal basis of these methods lies in Pareto front analysis, in which candidate solutions are ranked according to dominance relationships across all objectives. Any other candidate does not dominate molecules on the Pareto front and thus represents the best achievable compromise. By sampling and analyzing this front, researchers can rationally select materials tailored to specific application priorities—for example, prioritizing stability over marginal efficiency gains for long-term deployment.[128] This shift from single-metric optimization to trade-off-aware design

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

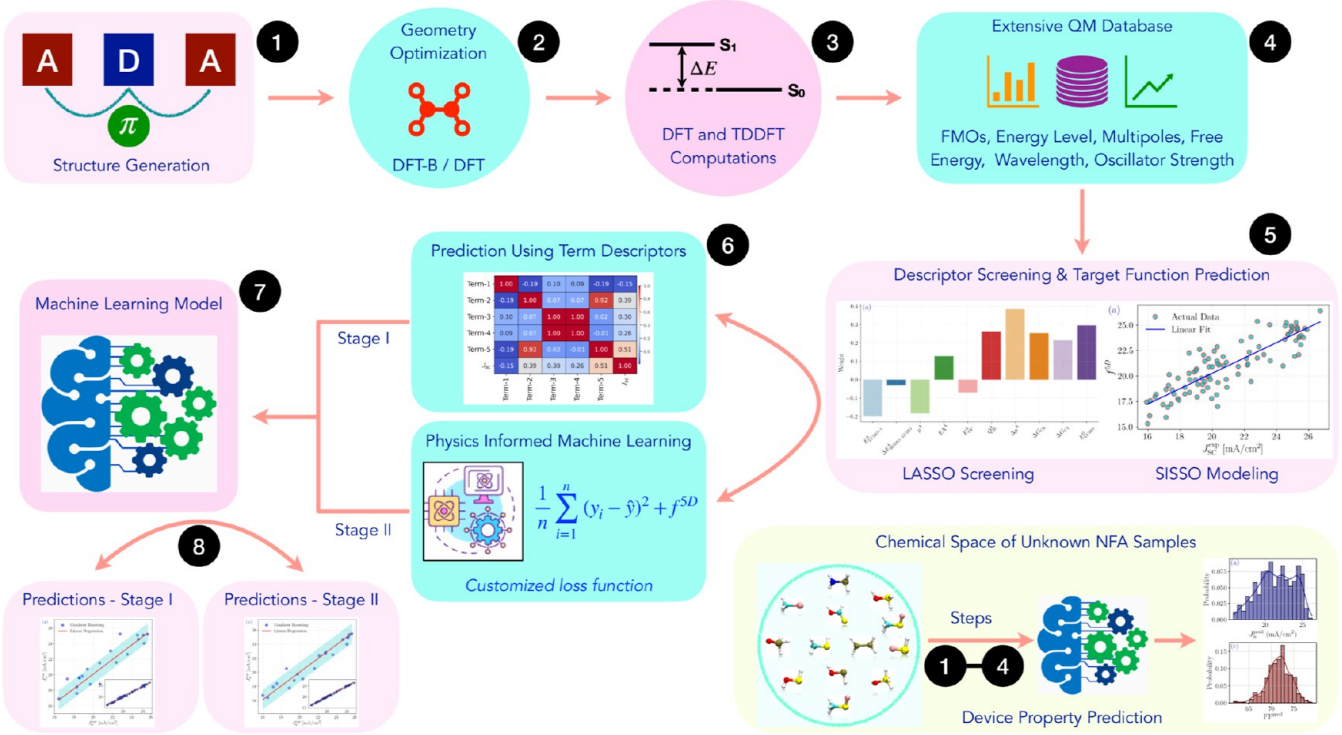


Figure 12. Schematic illustration of the physics-informed machine learning (PIML)-driven computational workflow employed for the high-throughput screening of active layer materials in organic solar cells, including data acquisition, descriptor extraction, physics-guided model training, prediction of photovoltaic performance metrics, and selection of promising candidate materials. Reprinted with permission from ref 61. Copyright 2025 American Chemical Society.

marks a critical maturation of generative machine learning for organic solar cells.

4.8. Promise and Pitfalls of Generative Molecular Design

4.8.1. Strengths. Generative machine learning models offer several compelling advantages for autonomous molecular design. First, they enable efficient exploration of vast chemical spaces—often comprising millions to billions of candidate molecules—far beyond the reach of manual design or conventional experimental screening. Second, multiple studies have demonstrated that generative models can propose molecules with experimentally validated high performance, underscoring their genuine predictive and design capability. Third, these models can be steered toward specific objectives using property-conditioned generation or reinforcement learning strategies, enabling targeted molecular discovery rather than blind exploration. Fourth, generative approaches can access regions of chemical space that are distant from historical training examples, thereby facilitating the discovery of structurally novel molecular classes with potentially superior properties. Finally, modern generative frameworks increasingly incorporate chemical rules and constraints, improving the likelihood that generated molecules are chemically valid, stable, and synthetically feasible.

4.8.2. Limitations. Despite their promise, several limitations currently restrict the practical impact of generative models. First, the availability of high-quality experimental data sets remains severely limited; for example, fewer than approximately 1000 nonfullerene acceptors (NFAs) have been experimentally characterized in detail, which constrains robust model training. Although transfer learning from large computational data sets partially alleviates this issue, such models may still fail to capture the subtle structure−property relationships governing experimental device performance. Second, generative models generally lack interpretability, offering limited chemical insight into why specific molecular features lead to improved properties. Third, computational efficiency remains a bottleneck—generating and screening millions of candidates using quantum chemical calculations or surrogate ML models can be prohibitively expensive. Fourth, a persistent gap exists between predicted and experimentally realized performance, with many high-performing in silico candidates failing during experimental validation. Finally, ensuring that generated molecules are not only chemically valid but also synthetically accessible and operationally stable remains challenging; many novel candidates prove difficult to synthesize or unstable under realistic device conditions.

5. METHODOLOGICAL ADVANCES AND CRITICAL PERSPECTIVES

5.1. Feature Engineering for OSC Performance Prediction

The choice of molecular descriptors fundamentally governs the performance of machine learning models. Early studies relied on relatively simple descriptors capturing frontier orbital energies, optical band gaps, and basic molecular size metrics. While informative, such minimal descriptor sets proved insufficient to capture the full complexity of OSC physics. More recent approaches have therefore incorporated richer descriptor spaces, including molecular polarity (dipole and quadrupole moments), charge distribution characteristics, reorganization energies relevant to charge transport, and descriptors capturing absorption properties such as extinction coefficients and oscillator strengths. Across multiple studies, several descriptor classes consistently emerge as dominant contributors to OSC performance prediction. Frontier orbital energies (HOMO and

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

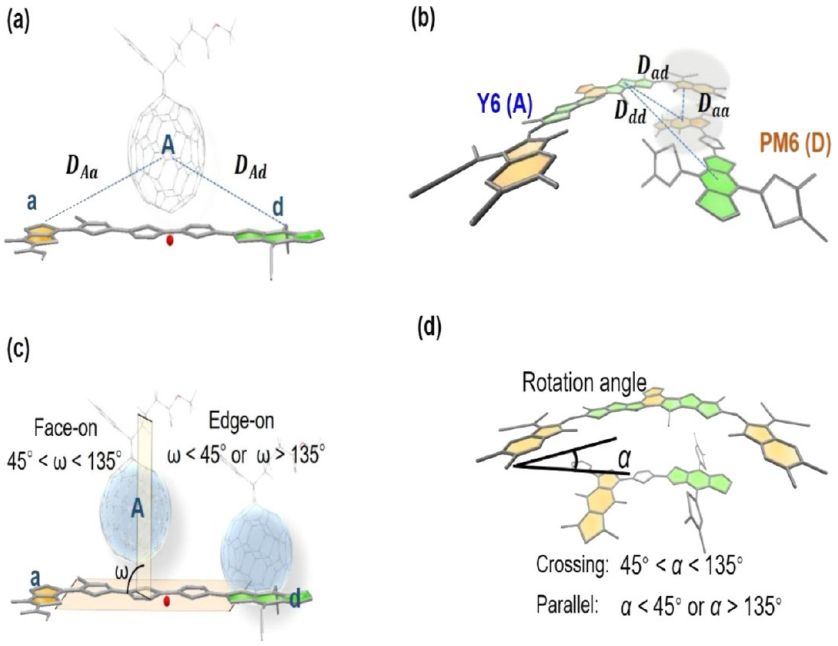


Figure 13. Definitions of geometric descriptors for intermolecular configurations in organic photovoltaic materials (a) effective distances D_{Aa} and D_{Ad} for fullerene systems, characterizing donor−acceptor spatial separation; (b) minimum intermolecular distances D_{aa}^{min}, D_{dd}^{min} and D_{ad}^{min} for nonfullerene systems; (c) inclination angle ω, defining the tilt between molecular planes; (d) rotation angle α, quantifying relative torsion between interacting molecular units. Reprinted with permission from ref 69. Copyright 2025 Springer Nature and the authors.

LUMO) remain among the most influential features, reflecting their central role in band alignment and driving forces for charge transfer. Optical band gaps and extinction coefficients are particularly important for predicting J_{SC}, as they directly encode light-harvesting capability. Notably, molecular quadrupole moments—which capture anisotropic charge distributions—have emerged as key descriptors for V_{OC} prediction, owing to their influence on intermolecular electrostatic interactions and charge-transfer state energetics. The prominence of quadrupole moments is especially significant, as this descriptor was largely underappreciated in traditional OSC material design. Recent computational studies systematically probing quadrupole effects demonstrate that maintaining quadrupole moments within an optimal range (approximately 80−130 ea_0^2) balances competing photovoltaic factors and improves device efficiency. This insight, uncovered through machine-learning-based feature importance analysis, provides concrete, actionable design guidance for experimental chemists developing next-generation NFAs.[132,133]

5.2. Interpretable and Physics-Guided ML Frameworks

A fundamental limitation of conventional machine learning lies in its lack of intrinsic physical interpretability and its susceptibility to predictions that violate basic physical constraints. When extrapolated beyond the training distribution, unconstrained ML models may predict open-circuit voltages exceeding theoretical limits imposed by donor−acceptor band offsets, or short-circuit currents exceeding limits set by light absorption.

Physics-Informed Machine Learning (PIML) frameworks address these shortcomings by explicitly embedding physical principles into the learning process. The Sure Independence Screening and Sparsifying Operator (SISSO) algorithm exemplifies this approach by identifying compact, human-readable mathematical expressions linking quantum-mechanical descriptors to photovoltaic performance metrics (see Figure 12). These data-driven equations, constructed through automated feature selection from large descriptor pools, explicitly reveal how quantities such as frontier orbital energies, charge-transfer energies, dipole differences, and quadrupole moments collectively govern voltage output and charge-transport efficiency. Recent efforts combining SISSO-derived relationships with gradient-boosting models have led to hybrid PIML architectures in which physically meaningful equations serve as constraints within ML loss functions. This strategy retains the predictive strength of flexible ML models while anchoring predictions in explicit physical relationships. The resulting frameworks simultaneously deliver high accuracy and interpretability, representing a critical step toward ML models that are both reliable and scientifically transparent.[61]

5.3. Linking Molecular Structure, Morphology, and Performance

One of the most formidable challenges in OSC modeling is the role of morphology—the nanoscale organization of donor and acceptor domains, crystallinity, domain size, and interfacial structure. While molecular-level electronic descriptors provide valuable insights, device performance depends critically on how materials organize in the solid state. Charge transport, exciton dissociation, and recombination losses are strongly governed by morphology, an emergent feature notoriously challenging to predict directly from isolated molecular properties.[36,37,39−41,61,106−110,134]

Several strategies have been explored to incorporate morphological information into ML models. The direct inclusion of experimentally measured morphological parameters—obtained from techniques such as atomic force microscopy, transmission electron microscopy, or grazing-incidence X-ray diffraction—substantially improves prediction accuracy. However, this approach inherently requires experimental characterization and therefore cannot eliminate experimental bottlenecks.[70,111,135,136] Molecular dynamics simulations offer an alternative route by predicting morphology from molecular

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

structure,[64,112,137−140] enabling the extraction of morphology-derived descriptors for ML input. Nevertheless, MD simulations of bulk heterojunction systems are computationally intensive, often requiring days per material, which limits scalability for high-throughput screening. A third approach involves ML models trained to predict morphology directly from molecular structure. However, current accuracy remains lower than that achieved for electronic property prediction, underscoring the emergent and collective nature of morphological phenomena. Recent studies on NFA-based OSCs clearly demonstrate that device efficiency is dictated not only by frontier orbital energetics but also—critically—by morphology-driven intermolecular packing and donor−acceptor miscibility (see Figure 13). Integrated ML, molecular dynamics, and TD-DFT analyses reveal that high-efficiency systems such as PM6/Y6 and ITIC- or Y6-derived blends exhibit favorable $\pi-\pi$ stacking distances, mixed face-on and edge-on packing motifs, and controlled aggregation behavior. Optimized donor−acceptor ratios enhance backbone-mediated packing continuity, reduce energetic disorder, and suppress nonradiative recombination. Collectively, these findings highlight that rational control of mesoscale morphology—via molecular design and processing conditions—offers a powerful lever for improving exciton dissociation and charge extraction, extending beyond traditional fullerene-based design paradigms.[69]

5.4. Generalization, Extrapolation, and Distribution Shift

Machine learning models typically achieve strong predictive performance when evaluated on test data drawn from the same distribution as the training set. However, accuracy degrades sharply when models are applied to novel chemical spaces—a phenomenon known as distribution shift. This challenge is particularly acute in materials discovery, where the most promising candidates often lie far outside the domain of existing data.[101−105]

Several strategies have been proposed to mitigate extrapolation failure. Uncertainty quantification methods, which associate confidence estimates with predictions, allow researchers to prioritize candidates with reliable predictions rather than treating all high-performance predictions equivalently.[79] Bayesian models, ensemble learning approaches, and evidential learning frameworks provide practical routes to uncertainty estimation.[65,81,90,141,142] Domain adaptation techniques from transfer learning can further reduce distribution shift when models trained on one class of materials are applied to chemically distinct systems.[143−145] Synthetic data generation using generative models offers another avenue for enriching training distributions in targeted regions of chemical space.[59,146] Finally, active learning strategies—in which models iteratively identify the most informative candidates for experimental validation—enable efficient expansion of training data sets toward regions of highest scientific interest.[59,101,141,147−149]

6. BENCHMARKING HIGH-EFFICIENCY ORGANIC SOLAR CELLS AND NFA DEVELOPMENT

6.1. Landmark Experimental Advances in OSC Efficiency

To place machine-learning-driven advances in a proper context, it is essential to first review the landmark experimental breakthroughs that have defined the current efficiency landscape of organic solar cells. Over the past decade, sustained progress in molecular design, morphology control, and device engineering has driven single-junction OSC efficiencies toward and beyond 19%, establishing stringent benchmarks for computational discovery efforts. Cui et al.[1] reported power conversion efficiencies approaching 18% in single-junction organic photovoltaic devices through precise molecular engineering combined with optimized device architectures, marking a pivotal milestone in OSC performance. Building on this progress, Gao et al.[17] introduced a donor alloy strategy that enables fine-tuning of the charge-transfer state energetics, achieving efficiencies exceeding 19.2%. This work clearly demonstrated that deliberate molecular-level manipulation of donor components can decisively influence photovoltaic performance. Subsequent studies highlighted the equally critical role of acceptor morphology in determining device efficiency. Li et al.[19] showed that controlled fibrillization of nonfullerene acceptors significantly improves nanoscale morphology and charge transport, yielding state-of-the-art efficiencies of approximately 19% in pseudo-bulk heterojunction devices. This study underscored that electronic optimization alone is insufficient and that morphology engineering is a key complementary design lever. Along similar lines, Liu et al.[21] developed a novel two-dimensionally conjugated nonfullerene acceptor, achieving efficiencies exceeding 19% and validating that extending conjugation dimensionality can simultaneously enhance absorption, packing, and charge transport. Further innovation in charge-transport engineering was demonstrated by Zhan et al.,[22] who introduced intermediary electron-acceptor channels to facilitate more efficient charge extraction. This strategy achieved record efficiencies of up to 19.3%, underscoring that optimizing charge-transfer pathways at the molecular and mesoscale levels provides a powerful route to further performance gains.

Collectively, these experimental milestones establish the performance benchmarks that contemporary machine learning and generative design strategies seek to reproduce and surpass. They also define the multifaceted design space—encompassing molecular structure, morphology, and charge-transport pathways—that data-driven approaches must capture to deliver experimentally relevant predictions.[1,17,19,21,22]

6.2. Evolution of Acceptor Materials and Structure−Property Design Rules

Understanding how acceptor molecular structure governs device performance has evolved through a combination of systematic experimental investigations and complementary computational analyses. Early structure−property studies revealed that subtle variations in acceptor architecture can exert disproportionate effects on charge transport, morphology, and ultimately device efficiency. For instance, Han et al. demonstrated that terminal stacking motifs critically control molecular packing and electron mobility in A−A type nonfullerene acceptors, establishing a direct link between molecular design and solid-state transport.[150] In related work, the same group showed that fluorination and π-extension of ITIC-based end groups enhance electron mobility by promoting tighter packing and improved intermolecular interactions.[151] Together, these studies underscore that even minor structural modifications—particularly at terminal units—can substantially reshape electronic coupling and charge-transport pathways.

The broader transition from fullerene acceptors to nonfullerene acceptors fundamentally reshaped acceptor design strategies in OSCs. Fullerene derivatives initially dominated the field due to their favorable three-dimensional electron transport characteristics; however, their limited chemical tunability, weak near-infrared absorption, and tendency toward excessive aggregation imposed intrinsic efficiency ceilings. Mi et al.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

(2014) provided a comprehensive experimental analysis of fullerene derivatives, highlighting both their strengths in facilitating charge transport and the structural limitations that constrained further performance improvements.[152]

A decisive shift occurred with the emergence of nonfullerene acceptors. Nielsen et al. pioneered early NFA designs that outperformed fullerene-based systems, demonstrating the feasibility of replacing fullerenes with chemically tunable alternatives. This breakthrough catalyzed rapid exploration of new acceptor motifs.[153] Subsequent experimental efforts, including those by Lin et al., showed that NFAs could deliver competitive or superior efficiencies while offering improved optical absorption and morphological control.[24] Bai et al. further validated the concept by synthesizing IDT-IC-based acceptors, establishing fullerene-free OSCs as a viable and scalable platform.[154] As NFA design matured, research efforts expanded beyond fused-ring architectures. Ma et al. demonstrated that nonfused, three-dimensionally structured acceptors can achieve efficiencies approaching 16% while offering enhanced stability, directly challenging the prevailing assumption that extensive ring fusion is a prerequisite for high performance.[26] These findings broadened the design space for NFAs and highlighted alternative routes to balancing efficiency and operational stability. Comprehensive reviews by Wadsworth et al. and Luo et al. synthesized these advances, documenting the rapid evolution of NFA molecular design strategies that propelled OSC efficiencies beyond 15% and, more recently, beyond 18%.[35,135] Collectively, these experimental studies established a set of empirical structure−property design rules—linking conjugation length, end-group chemistry, molecular packing, and aggregation control—that continue to guide both experimental synthesis and machine-learning-driven materials discovery.

6.3. Closing the Loop: ML-Guided Design and Experimental Realization

Recent progress in OSC research demonstrates that machine learning models, when tightly integrated with generative design and experimental validation, can substantially accelerate the discovery of high-efficiency materials. Early evidence for this paradigm was provided by Wu et al.,[64] who developed ML models trained on curated donor−acceptor (D/A) data sets to predict power conversion efficiency in nonfullerene OSCs. Virtual screening identified high-performing candidates, several of which were subsequently synthesized and fabricated into devices, with measured efficiencies closely matching model predictions. This work provided one of the first clear demonstrations that ML-guided screening can directly translate into experimentally observed performance gains. Building on predictive screening, Tan et al.[127] advanced a fully integrated generative-predictive workflow for the discovery of NFAs. Their pipeline combined generative ML models with property predictors to propose entirely new acceptor structures optimized for high efficiency. Selected candidates were synthesized and incorporated into OSC devices, where their experimentally measured PCEs confirmed the validity of the ML-driven design strategy. This study demonstrated that ML frameworks can move beyond prioritizing known chemical space to autonomously proposing novel molecular architectures and validating their performance experimentally, effectively closing the design-test loop. Recent deep learning approaches further strengthened this integration. Sun and co-workers[59] introduced DeepAcceptor, a graph-based representation-learning framework that employs transformer architectures to predict PCE for small-molecule acceptors. Trained on a combination of quantum-computed and experimental data, the model achieved competitive predictive accuracy ($R^2 \approx 0.67$, MAE ≈ 1.78) and successfully guided the synthesis and experimental validation of top-ranked candidates, confirming its practical utility for acceptor discovery.

In parallel, Suthar et al.[56] assembled an extensive data set of experimentally characterized polymer/NFA systems and developed supervised ML models incorporating frontier orbital descriptors and RDKit-derived structural features. Their random forest models yielded robust PCE predictions (Pearson's $r \approx 0.79$, mean absolute percentage error ~2%), while gradient boosting further improved the prediction of individual device metrics such as J_{SC} and V_{OC}. Crucially, SHAP-based interpretability analysis identified key descriptors governing performance, enabling chemically intuitive, targeted optimization strategies rather than purely black-box predictions. Collectively, these studies demonstrate that ML-guided discovery pipelines—spanning predictive modeling, generative design, and experimental validation—can reliably identify high-efficiency OSC materials whose measured device performance closely tracks in silico expectations. This convergence of computation and experiment marks a decisive shift away from trial-and-error materials development toward data-driven, closed-loop optimization frameworks.[56,59,64,127]

7. PRACTICAL CONSTRAINTS AND MULTI-OBJECTIVE DESIGN IN ML-GUIDED OSC DISCOVERY

Despite notable progress in machine-learning-assisted discovery of organic photovoltaic materials, several interconnected challenges limit the translation of high-performing computational predictions into experimentally viable devices. These challenges span fundamental gaps between prediction and experiment, data limitations, chemical feasibility, and the inherent tension between model accuracy and interpretability. Addressing these issues naturally motivates a shift from single-objective efficiency optimization toward balanced, multi-objective molecular design.[128]

7.1. Bridging the Prediction-Experiment Divide

A persistent challenge in the field is the frequent mismatch between predicted high efficiencies and experimentally realized device performance. While successful prospective validations—achieving agreement within 1−2% PCE—have been reported, many computationally promising candidates fail during experimental implementation. This discrepancy arises from multiple compounding factors. Most ML models rely almost exclusively on molecular electronic descriptors, implicitly assuming that favorable intrinsic properties translate directly into high device performance. In practice, mesoscale morphology plays a decisive role: solid-state packing, phase separation, and interfacial structure critically influence charge transport and recombination. These morphological outcomes depend sensitively on processing conditions—such as solvent choice, annealing protocols, additives, and device architecture—that are rarely incorporated into predictive frameworks.

The persistence of this gap should therefore be understood primarily as a data-fidelity and problem-definition issue rather than a simple limitation of model architecture. Most retrospective ML models are trained on literature-level summaries of donor and acceptor identity, selected electronic descriptors, and headline device metrics, whereas experimental

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

realization depends on processing-controlled morphology, interfacial microstructure, electrode and interlayer selection, thickness control, and synthetic feasibility. As a result, a candidate may appear highly promising at the level of composition-based prediction yet fail experimentally because it does not reproduce the nanoscale organization or recombination landscape required to achieve the predicted PCE. This limitation becomes particularly pronounced when models are transferred from low-fidelity computational data sets to experimental systems, or when proposed candidates fall outside the structural and processing distributions represented in the training data.[59,64,65,69,101,105,127] In addition, the physics governing OSC operation remains incompletely resolved. Nonradiative recombination pathways, charge-transfer state energetics, and voltage loss mechanisms are still under active investigation. Consequently, ML models encode current understanding and may systematically miss governing factors not captured by available descriptors, limiting robustness when extrapolating beyond known material classes.[101−105]

7.2. Data Bias, Descriptor Fidelity, and Model Reliability

The reliability of ML predictions is fundamentally constrained by the quality and representativeness of training data. Current OSC data sets comprise on the order of 10^3 to 10^4 curated examples—substantial by chemical standards, yet small relative to descriptor dimensionality.[56,59,62] This imbalance increases susceptibility to overfitting and reduces confidence in generalization. Systematic data set bias further complicates model behavior. Experimental research naturally focuses on chemically intuitive or historically successful motifs, leading to the overrepresentation of specific structural classes and the underrepresentation of unconventional chemistries. Models trained on such data may preferentially favor familiar molecular patterns, artificially inflating predicted performance while penalizing structurally distinct candidates. Descriptor fidelity introduces an additional layer of uncertainty. Quantum chemical calculations used to generate electronic features carry systematic errors—most notably, bandgap underestimation in commonly employed DFT functionals such as B3LYP. While ML models often learn relative trends, absolute errors and functional-dependent shifts can degrade predictive reliability when extrapolating to unexplored regions of chemical space.[36,37,39−41,60,79,105,134]

7.3. Chemical Feasibility and Synthetic Constraints

Beyond predictive accuracy, chemical validity and experimental tractability impose practical constraints on ML-guided discovery. Generative models frequently propose molecules that are formally valid yet synthetically unrealistic, involving unstable motifs, impractical reaction pathways, or excessive synthetic complexity. Recent approaches mitigate these issues through fragment-based generation, incorporation of synthetic accessibility scores, and staged filtering pipelines. Human-in-the-loop strategies—where expert chemists evaluate top-ranked candidates—remain essential for ensuring feasibility, though they introduce subjectivity and limit throughput.[127] In practice, the most robust workflows combine modular molecular construction, cheminformatic screening, computational feasibility metrics, and expert review, balancing novelty with experimental realism.[56,59,64,127]

7.4. Interpretability, Trust, and Physics Awareness

The increasing complexity of ML models introduces a trade-off between predictive power and interpretability. Linear and shallow models offer transparency but often fail to capture nonlinear structure−property relationships, whereas deep learning approaches achieve superior accuracy at the cost of explainability. This opacity raises legitimate concerns regarding trust and physical insight. Interpretability tools such as SHAP analysis, attention mechanisms, and dimensionality reduction provide partial remedies, enabling posthoc rationalization of predictions and identification of dominant features. Physics-informed approaches, including SISSO-derived models and constraint-embedded learning, offer a more direct link between prediction and mechanism.[61,82] However, interpretability remains imperfect, and explanatory frameworks do not always reflect true causality.

7.5. From Single-Metric Optimization to Balanced Molecular Design

These limitations collectively highlight the inadequacy of optimizing PCE alone.[105,115,129] Commercially viable OSC materials must balance efficiency with stability, synthetic accessibility, cost, and environmental considerations. Multi-objective optimization frameworks provide a systematic means of navigating these trade-offs.[128,131] Genetic algorithms and related evolutionary strategies are particularly well suited for this task, naturally exploring Pareto-optimal solutions across competing objectives. When coupled with ML surrogate models for rapid property evaluation, these approaches enable efficient traversal of chemical space while maintaining diversity and avoiding premature convergence. Overall, progress in ML-guided OSC discovery increasingly depends not on marginal gains in predictive accuracy, but on integrating physical realism, chemical feasibility, and multiobjective reasoning into coherent, experimentally grounded design pipelines.

8. FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

Continued progress in machine-learning-guided OSC discovery will depend on advancing beyond purely data-driven prediction toward frameworks that integrate physical constraints, first-principles accuracy, multiscale understanding, and experimental validation. Several complementary research directions offer clear opportunities to improve both predictive reliability and practical impact.

Physics-informed and constraint-aware learning represents a particularly promising avenue. Embedding physical laws, thermodynamic limits, and conservation principles directly into learning architectures can ensure that predictions remain physically meaningful, especially when models extrapolate beyond their training distributions. In practice, this may involve augmenting loss functions with penalties for violating known bounds (e.g., predicted V_{OC} exceeding bandgap-determined limits), incorporating analytic thermodynamic or kinetic relations into network architectures, or employing adversarial schemes in which a physics-aware critic enforces consistency. Integration of SISSO-derived symbolic relationships into neural network training—either as constraints or regularization terms—offers an attractive route to combine interpretability with nonlinear modeling capacity.

Tighter integration between machine learning and first-principles quantum chemistry is equally important. Rather than serving as alternatives, these approaches are naturally complementary within active learning frameworks. Starting from limited experimental or computational data sets, ML models can identify promising candidates and regions of high

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

uncertainty, guiding targeted quantum chemical calculations or experimental synthesis. Newly generated data are then incorporated to refine models, iteratively improving accuracy and generalizability while minimizing computational and experimental cost. Such closed-loop strategies concentrate effort where it is most informative, accelerating convergence toward reliable predictive models.

Multiscale and multiphysics modeling remains a critical gap in current workflows. Most ML models directly map molecular structure to device-level metrics, implicitly compressing a complex sequence of physical processes into a single step. In reality, device performance emerges from coupled phenomena spanning electronic structure, nanoscale morphology, mesoscale organization, and device architecture. Hierarchical frameworks that link molecular descriptors to morphology prediction and, subsequently, to device-scale simulations offer a more physically grounded pathway from chemical design to measured performance. Although computationally demanding, such approaches promise greater robustness and interpretability than direct end-to-end prediction.

Expanding structural diversity and addressing distribution shift will be essential for sustained discovery. Existing models are heavily trained on well-established NFA families, limiting their reliability for emerging material classes such as nonfused acceptors, ladder-type systems, or novel core scaffolds. Deliberate inclusion of structurally diverse and less-characterized materials—combined with transfer learning and domain adaptation techniques—can improve generalization and reduce overfitting to familiar motifs. Purposeful exploration of atypical or out-of-distribution candidates may also reveal promising design directions that conventional intuition overlooks.

Experimental automation and high-throughput validation will ultimately determine the practical impact of ML-guided discovery. Computational screening is now orders of magnitude faster than experimental validation, creating a bottleneck at the synthesis and device-fabrication stage. Advances in robotic synthesis, automated processing, and high-throughput characterization provide a pathway to address this imbalance. Integrating ML-driven candidate selection with automated experimental platforms enables closed-loop discovery cycles, in which computational predictions, rapid synthesis, and characterization iteratively inform one another, substantially accelerating materials development.

Finally, cross-domain learning and knowledge transfer represent an underexploited opportunity. OSCs share conceptual and methodological challenges with related materials classes, including perovskite solar cells, organic light-emitting diodes, organic transistors, photocatalysts, and battery materials. Transfer learning across these domains can leverage shared descriptors, design principles, and optimization strategies, reducing data requirements and improving model robustness. Developing generalizable materials, informatics frameworks, and shared descriptor vocabularies will be key to enabling such cross-domain synergy. Together, these directions point toward a more integrated, physically grounded, and experimentally connected paradigm for machine-learning-assisted OSC discovery—setting the stage for the concluding perspective.

9. CONCLUSION

The integration of machine learning and artificial intelligence into organic solar cell research marks a substantive shift in how materials discovery and device optimization are pursued. By moving beyond predominantly Edisonian trial-and-error experimentation toward data-driven, predictive design, the field has gained powerful tools to accelerate the identification of high-efficiency, potentially commercially viable organic photovoltaic materials.

Predictive ML models—ranging from classical algorithms such as random forests and gradient boosting to advanced deep learning architectures including graph neural networks and transformers—have demonstrated a genuine ability to correlate molecular structure with device-level performance. Importantly, multiple prospective validation studies have confirmed that computationally identified materials can translate into experimentally realized devices with efficiencies close to the predicted values. With current state-of-the-art models achieving mean absolute errors on the order of 1−2 percentage points in PCE, these approaches are now sufficiently accurate to meaningfully guide experimental prioritization.

Generative machine learning methods further extend this capability by enabling inverse design. Approaches such as variational autoencoders, generative adversarial networks, genetic algorithms, and reinforcement learning have demonstrated that large regions of chemical space can be efficiently explored, yielding structurally novel candidate materials with high predicted performance. Experimental validation of generatively proposed acceptors demonstrates that computational design is no longer limited to screening known chemistries but can actively propose viable new molecular architectures.

Despite this progress, important limitations remain. The persistent gap between predicted and realized performance reflects the difficulty of capturing all relevant physical phenomena within current models. Morphology formation, nonradiative recombination pathways, and processing-dependent effects are only weakly constrained by molecular descriptors alone. Chemical validity and synthetic accessibility remain challenging for de novo generative models, and distribution shift continues to limit predictive reliability when extrapolating to unfamiliar regions of chemical space.

Looking forward, progress will likely arise from the convergence of several complementary strategies. Physics-informed machine learning offers a path toward improved robustness and interpretability by embedding fundamental constraints directly into learning frameworks. Active learning approaches integrating ML with first-principles quantum chemistry can balance computational efficiency with physical accuracy. Multiscale modeling frameworks promise tighter connections between molecular design, morphology, and device-level performance. Finally, advances in experimental automation and closed-loop computational−experimental workflows will be essential for translating computational predictions into rapid, reliable validation.

Taken together, the field is approaching a critical inflection point. Machine learning has matured from a purely exploratory tool into a practical component of the OSC discovery pipeline, capable of accelerating experimentation and expanding accessible chemical space. As methodologies continue to improve, ML is poised not to replace chemical intuition or experimental expertise, but to amplify them—enabling more systematic exploration, more informed decision-making, and faster progress toward next-generation organic photovoltaic technologies.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Data Availability Statement

No new data sets were generated during the current study. All computational and experimental data discussed are derived from previously published studies cited in this review.

■ AUTHOR INFORMATION

Corresponding Author

Anirban Mondal − *Department of Chemistry, Indian Institute of Technology Gandhinagar, Gandhinagar 382355 Gujarat, India;* orcid.org/0000-0003-3029-8840; Email: amondal@iitgn.ac.in

Author

Bibhas Das − *Department of Chemistry, Indian Institute of Technology Gandhinagar, Gandhinagar 382355 Gujarat, India;* orcid.org/0000-0002-9671-5275

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.6c01194

Author Contributions

AM conceived the scope and structure of the review. BD conducted the literature survey and computational analyses. AM and BD jointly analyzed the results, discussed the interpretation of the findings, and cowrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the Indian Institute of Technology Gandhinagar, India, for providing research facilities and financial support. B.D. and A.M. thank PARAM Ananta for computational resources.

■ REFERENCES

(1) Cui, Y.; Yao, H.; Zhang, J.; Xian, K.; Zhang, T.; Hong, L.; Wang, Y.; Xu, Y.; Ma, K.; An, C.; He, C.; Wei, Z.; Gao, F.; Hou, J. Single-junction organic photovoltaic cells with approaching 18% efficiency. *Adv. Mater.* **2020**, *32*, 1908205.

(2) Cartlidge, E. Bright outlook for solar cells. *Phys. World* **2007**, *20*, 20.

(3) Maka, A. O.; Alabid, J. M. Solar energy technology and its roles in sustainable development. *Clean Energy* **2022**, *6*, 476−483.

(4) Novas, N.; Garcia, R. M.; Camacho, J. M.; Alcayde, A. Advances in solar energy towards efficient and sustainable energy. *Sustainability* **2021**, *13*, 6295.

(5) Ganguly, G. Improved sustainability of solar panels by improving stability of amorphous silicon solar cells. *Sci. Rep.* **2023**, *13*, 10512.

(6) Yu, Z.-P.; Liu, Z.-X.; Chen, F.-X.; Qin, R.; Lau, T.-K.; Yin, J.-L.; Kong, X.; Lu, X.; Shi, M.; Li, C.-Z.; Chen, H. Simple non-fused electron acceptors for efficient and stable organic solar cells. *Nat. Commun.* **2019**, *10*, 2152.

(7) Kalowekamo, J.; Baker, E. Estimating the manufacturing cost of purely organic solar cells. *Sol. Energy* **2009**, *83*, 1224−1231.

(8) Batista, D.; Oliveira, L. B.; Paulino, N.; Carvalho, C.; Oliveira, J. P.; Farinhas, J.; Charas, A.; Dos Santos, P. M. Combined organic photovoltaic cells and ultra low power CMOS circuit for indoor light energy harvesting. *Sensors* **2019**, *19*, 1803.

(9) Zhu, J.; Xia, J.; Li, Y.; Li, Y. Perspective on flexible organic solar cells for self-powered wearable applications. *ACS Appl. Mater. Interfaces* **2025**, *17*, 5595−5608.

(10) Ali, A. O.; Elgohr, A. T.; El-Mahdy, M. H.; Zohir, H. M.; Emam, A. Z.; Mostafa, M. G.; Al-Razgan, M.; Kasem, H. M.; Elhadidy, M. S. Advancements in photovoltaic technology: A comprehensive review of recent advances and future prospects. *Energy Convers. Manage.:X* **2025**, *26*, 100952.

(11) Yoon, S.; Park, S.; Park, S. H.; Nah, S.; Lee, S.; Lee, J.-W.; Ahn, H.; Yu, H.; Shin, E.-Y.; Kim, B. J.; Min, B. K.; Noh, J. H.; Son, H. J. High-performance scalable organic photovoltaics with high thickness tolerance from 1 cm^2 to above 50 cm^2. *Joule* **2022**, *6*, 2406−2422.

(12) Liu, S.; Wang, H.; Cui, Y.; Zeng, S.; Sun, C.; Li, H.; Li, H.; Ye, L.; Yuan, H.; Zhu, H.; Yu, J.; Chen, H.; Hu, X.; Chen, Y. Sustainable eco-friendly printing of high-performance large-area organic photovoltaics via enhanced Laplace pressure gradient. *Nat. Commun.* **2025**, *16*, 8520.

(13) Wadsworth, A.; Hamid, Z.; Kosco, J.; Gasparini, N.; McCulloch, I. The bulk heterojunction in organic photovoltaic, photodetector, and photocatalytic applications. *Adv. Mater.* **2020**, *32*, 2001763.

(14) Solak, E. K.; Irmak, E. Advances in organic photovoltaic cells: a comprehensive review of materials, technologies, and performance. *RSC Adv.* **2023**, *13*, 12244−12269.

(15) Han, G.; Guo, Y.; Song, X.; Wang, Y.; Yi, Y. Terminal − stacking determines three-dimensional molecular packing and isotropic charge transport in an A − A electron acceptor for non-fullerene organic solar cells. *J. Mater. Chem. C* **2017**, *5*, 4852−4857.

(16) Han, G.; Guo, Y.; Ning, L.; Yi, Y. Improving the Electron Mobility of ITIC by End-Group Modulation: The Role of Fluorination and -Extension. *Sol. RRL* **2019**, *3*, 1800251.

(17) Gao, J.; Yu, N.; Chen, Z.; Wei, Y.; Li, C.; Liu, T.; Gu, X.; Zhang, J.; Wei, Z.; Tang, Z.; Hao, X.; Zhang, F.; Zhang, X.; Huang, H. Over 19.2% efficiency of organic solar cells enabled by precisely tuning the charge transfer state via donor alloy strategy. *Adv. Sci.* **2022**, *9*, 2203606.

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

(19) Li, D.; Deng, N.; Fu, Y.; Guo, C.; Zhou, B.; Wang, L.; Zhou, J.; Liu, D.; Li, W.; Wang, K.; Sun, Y.; Wang, T. Fibrillization of Non-Fullerene Acceptors Enables 19% Efficiency Pseudo-Bulk Heterojunction Organic Solar Cells. *Adv. Mater.* **2023**, *35*, 2208211.

(20) Chen, H.; Jeong, S. Y.; Tian, J.; Zhang, Y.; Naphade, D. R.; Alsufyani, M.; Zhang, W.; Griggs, S.; Hu, H.; Barlow, S.; Woo, H. Y.; Marder, S. R.; Anthopoulos, T. D.; McCulloch, I.; Lin, Y. A 19% efficient and stable organic photovoltaic device enabled by a guest nonfullerene acceptor with fibril-like morphology. *Energy Environ. Sci.* **2023**, *16*, 1062−1070.

(21) Liu, K.; Jiang, Y.; Liu, F.; Ran, G.; Huang, F.; Wang, W.; Zhang, W.; Zhang, C.; Hou, J.; Zhu, X. Organic Solar Cells with Over 19% Efficiency Enabled by a 2D-Conjugated Non-fullerene Acceptor Featuring Favorable Electronic and Aggregation Structures. *Adv. Mater.* **2023**, *35*, 2300363.

(22) Zhan, L.; Li, S.; Li, Y.; Sun, R.; Min, J.; Chen, Y.; Fang, J.; Ma, C.-Q.; Zhou, G.; Zhu, H.; Zuo, L.; Qiu, H.; Yin, S.; Chen, H. Manipulating charge transfer and transport via intermediary electron acceptor channels enables 19.3% efficiency organic photovoltaics. *Adv. Energy Mater.* **2022**, *12*, 2201076.

(23) Bi, P.; Zhang, S.; Ren, J.; Chen, Z.; Zheng, Z.; Cui, Y.; Wang, J.; Wang, S.; Zhang, T.; Li, J.; Xu, Y.; Qin, J.; An, C.; Ma, W.; Hao, X.; Hou, J. A high-performance nonfused wide-bandgap acceptor for versatile photovoltaic applications. *Adv. Mater.* **2022**, *34*, 2108090.

(24) Lin, Y.; Wang, J.; Zhang, Z.-G.; Bai, H.; Li, Y.; Zhu, D.; Zhan, X. An electron acceptor challenging fullerenes for efficient polymer solar cells. *Adv. Mater.* **2015**, *27*, 1170−1174.

(25) Bai, H.; Wang, Y.; Cheng, P.; Wang, J.; Wu, Y.; Hou, J.; Zhan, X. An electron acceptor based on indacenodithiophene and 1, 1-dicyanomethylene-3-indanone for fullerene-free organic solar cells. *J. Mater. Chem. A* **2015**, *3*, 1910−1914.

(26) Ma, L.; Zhang, S.; Zhu, J.; Wang, J.; Ren, J.; Zhang, J.; Hou, J. Completely non-fused electron acceptor with 3D-interpenetrated

crystalline structure enables efficient and stable organic solar cell. *Nat. Commun.* **2021**, *12*, 5093.
(27) Yao, H.; Cui, Y.; Qian, D.; Ponseca, C. S.; Honarfar, A.; Xu, Y.; Xin, J.; Chen, Z.; Hong, L.; Gao, B.; Yu, R.; Zu, Y.; Ma, W.; Chabera, P.; Pullerits, T.; Yartsev, A.; Gao, F.; Hou, J. 14.7% efficiency organic photovoltaic cells enabled by active materials with a large electrostatic potential difference. *J. Am. Chem. Soc.* **2019**, *141*, 7743−7750.
(28) Yuan, J.; Zhang, Y.; Zhou, L.; Zhang, G.; Yip, H.-L.; Lau, T.-K.; Lu, X.; Zhu, C.; Peng, H.; Johnson, P. A.; Leclerc, M.; Cao, Y.; Ulanski, J.; Li, Y.; Zou, Y. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core. *Joule* **2019**, *3*, 1140−1151.
(29) Han, J.; Xu, H.; Paleti, S. H. K.; Sharma, A.; Baran, D. Understanding photochemical degradation mechanisms in photoactive layer materials for organic solar cells. *Chem. Soc. Rev.* **2024**, *53*, 7426−7454.
(30) Duan, L.; Uddin, A. Progress in stability of organic solar cells. *Adv. Sci.* **2020**, *7*, 1903259.
(31) Huang, Y.; Kramer, E. J.; Heeger, A. J.; Bazan, G. C. Bulk heterojunction solar cells: morphology and performance relationships. *Chem. Rev.* **2014**, *114*, 7006−7043.
(32) Rosenthal, K. D.; Hughes, M. P.; Luginbuhl, B. R.; Ran, N. A.; Karki, A.; Ko, S.-J.; Hu, H.; Wang, M.; Ade, H.; Nguyen, T.-Q. Quantifying and understanding voltage losses due to nonradiative recombination in bulk heterojunction organic solar cells with low energetic offsets. *Adv. Energy Mater.* **2019**, *9*, 1901077.
(33) Heeger, A. J. 25th anniversary article: bulk heterojunction solar cells: understanding the mechanism of operation. *Adv. Mater.* **2014**, *26*, 10−28.
(34) Song, J.; Zhang, M.; Yuan, M.; Qian, Y.; Sun, Y.; Liu, F. Morphology characterization of bulk heterojunction solar cells. *Small Methods* **2018**, *2*, 1700229.
(35) Luo, D.; Jang, W.; Babu, D. D.; Kim, M. S.; Wang, D. H.; Kyaw, A. K. K. Recent progress in organic solar cells based on non-fullerene acceptors: materials to devices. *J. Mater. Chem. A* **2022**, *10*, 3255−3295.
(36) Mahmood, A.; Wang, J.-L. Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy Environ. Sci.* **2021**, *14*, 90−105.
(37) Rodríguez-Martínez, X.; Pascual-San-José, E.; Campoy-Quiles, M. Accelerating organic solar cell material's discovery: high-throughput screening and big data. *Energy Environ. Sci.* **2021**, *14*, 3301−3322.
(38) Zhang, G.; Lin, F. R.; Qi, F.; Heumüller, T.; Distler, A.; Egelhaaf, H.-J.; Li, N.; Chow, P. C. Y.; Brabec, C. J.; Jen, A. K.-Y.; Yip, H.-L. Renewed prospects for organic photovoltaics. *Chem. Rev.* **2022**, *122*, 14180−14274.
(39) Greenstein, B. L.; Hutchison, G. R. Organic photovoltaic efficiency predictor: data-driven models for non-fullerene acceptor organic solar cells. *J. Phys. Chem. Lett.* **2022**, *13*, 4235−4243.
(40) Sun, W.; Zheng, Y.; Zhang, Q.; Yang, K.; Chen, H.; Cho, Y.; Fu, J.; Odunmbaku, O.; Shah, A. A.; Xiao, Z.; Lu, S.; Chen, S.; Li, M.; Qin, B.; Yang, C.; Frauenheim, T.; Sun, K. Artificial Intelligence Designer for Highly-Efficient Organic Photovoltaic Materials. *J. Phys. Chem. Lett.* **2021**, *12*, 8847−8854.
(41) Miyake, Y.; Saeki, A. Machine learning-assisted development of organic solar cell materials: issues, analyses, and outlooks. *J. Phys. Chem. Lett.* **2021**, *12*, 12391−12401.
(42) Kim, G.-H.; Lee, C.; Kim, K.; Ko, D.-H. Novel structural feature-descriptor platform for machine learning to accelerate the development of organic photovoltaics. *Nano Energy* **2023**, *106*, 108108.
(43) Mahmood, A.; Irfan, A.; Wang, J.-L. 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.
(44) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(45) Islam, M. S.; Islam, M. T.; Sarker, S.; Jame, H. A.; Nishat, S. S.; Jani, M. R.; Rauf, A.; Ahsan, S.; Shorowordi, K. M.; Efstathiadis, H.; Carbonara, J.; Ahmed, S. Machine learning approach to delineate the impact of material properties on solar cell device Physics. *ACS Omega* **2022**, *7*, 22263−22278.
(46) Kranthiraja, K.; Saeki, A. Experiment-oriented machine learning of polymer: non-fullerene organic solar cells. *Adv. Funct. Mater.* **2021**, *31*, 2011168.
(47) Mahmood, A.; Irfan, A.; Wang, J.-L. Developing efficient small molecule acceptors with sp2-hybridized nitrogen at different positions by density functional theory calculations, molecular dynamics simulations and machine learning. *Chem.—Eur. J.* **2022**, *28*, No. e202103712.
(48) Mahmood, A.; Irfan, A.; Wang, J.-L. Machine learning for organic photovoltaic polymers: A minireview. *Chin. J. Polym. Sci.* **2022**, *40*, 870−876.
(49) Han, G.; Yi, Y. Singlet-triplet energy gap as a critical molecular descriptor for predicting organic photovoltaic efficiency. *Angew. Chem., Int. Ed.* **2022**, *49*, No. e202213953.
(50) Miyake, Y.; Kranthiraja, K.; Ishiwari, F.; Saeki, A. Improved Predictions of Organic Photovoltaic Performance through Machine Learning Models Empowered by Artificially Generated Failure Data. *Chem. Mater.* **2022**, *34*, 6912−6920.
(51) Zhao, Q.; Shan, Y.; Xiang, C.; Wang, J.; Zou, Y.; Zhang, G.; Liu, W. Predicting power conversion efficiency of binary organic solar cells based on Y6 acceptor by machine learning. *J. Energy Chem.* **2023**, *82*, 139−147.
(52) Zhao, Z.-W.; del Cueto, M.; Geng, Y.; Troisi, A. Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chem. Mater.* **2020**, *32*, 7777−7787.
(53) Suthar, R.; Abhijith, T.; Sharma, P.; Karak, S. Machine learning framework for the analysis and prediction of energy loss for non-fullerene organic solar cells. *Sol. Energy* **2023**, *250*, 119−127.
(54) Mahmood, A.; Sandali, Y.; Wang, J.-L. Easy and fast prediction of green solvents for small molecule donor-based organic solar cells through machine learning. *Phys. Chem. Chem. Phys.* **2023**, *25*, 10417−10426.
(55) Liu, C.; Lüer, L.; Corre, V. M. L.; Forberich, K.; Weitz, P.; Heumüller, T.; Du, X.; Wortmann, J.; Zhang, J.; Wagner, J.; Ying, L.; Hauch, J.; Li, N.; Brabec, C. J. Understanding Causalities in Organic Photovoltaics Device Degradation in a Machine-Learning-Driven High-Throughput Platform. *Adv. Mater.* **2023**, *36*, 2300259.
(56) Suthar, R.; Abhijith, T.; Karak, S. Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors. *J. Mater. Chem. A* **2023**, *11*, 22248−22258.
(57) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. 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.
(58) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry−the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698−704.
(59) Sun, J.; Li, D.; Zou, J.; Zhu, S.; Xu, C.; Zou, Y.; Zhang, Z.; Lu, H. Accelerating the discovery of acceptor materials for organic solar cells by deep learning. *npj Comput. Mater.* **2024**, *10*, 181.
(60) Das, B.; Mondal, A. Predictive Modeling and Design of Organic Solar Cells: A Data-Driven Approach for Material Innovation. *ACS Appl. Energy Mater.* **2024**, *7*, 9349−9363.
(61) Khatua, R.; Das, B.; Mondal, A. Physics-informed machine learning with data-driven equations for predicting organic solar cell performance. *ACS Appl. Mater. Interfaces* **2024**, *16*, 57467−57480.
(62) Liu, X.; Zhang, X.; Sheng, Y.; Zhang, Z.; Xiong, P.; Ju, X.; Zhu, J.; Ye, C. Advancing organic photovoltaic materials by machine learning-driven design with polymer-unit fingerprints. *npj Comput. Mater.* **2025**, *11*, 107.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

(63) Zhang, T.; Yuk Lin Lai, J.; Shi, M.; Li, Q.; Zhang, C.; Yan, H. Data cleansing and sub-unit-based molecular description enable accurate prediction of the energy levels of non-fullerene acceptors used in organic solar cells. *Adv. Sci.* **2024**, *11*, 2308652.
(64) Wu, Y.; Guo, J.; Sun, R.; Min, J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Comput. Mater.* **2020**, *6*, 120.
(65) Yang, H.; Wold, A.; Ou, J.; Rech, J. J.; You, W.; Wang, Y. OSC-Net: a multi-fidelity machine learning model for organic solar cells. *J. Mater. Chem. A* **2026**, *14*, 1208−1220.
(66) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(67) Zhang, C.-R.; Cao, R.; Liu, X.-M.; Zhang, M.-L.; Gong, J.-J.; Liu, Z.-J.; Wu, Y.-Z.; Chen, H.-S. Designing Donors and Nonfullerene Acceptors for Organic Solar Cells Assisted by Machine Learning and Fragment-Based Molecular Fingerprints. *Sol. RRL* **2025**, *9*, 2400846.
(68) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun, K. Machine learning−assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **2019**, *5*, No. eaay4275.
(69) Fu, L.; Hu, H.; Zhu, Q.; Zheng, L.; Gu, Y.; Wen, Y.; Ma, H.; Yin, H.; Ma, J. Machine learning assisted prediction of charge transfer properties in organic solar cells by using morphology-related descriptors. *Nano Res.* **2023**, *16*, 3588−3596.
(70) Kobayashi, Y.; Miyake, Y.; Ishiwari, F.; Ishiwata, S.; Saeki, A. Machine learning of atomic force microscopy images of organic solar cells. *RSC Adv.* **2023**, *13*, 15107−15113.
(71) Seifrid, M.; Lo, S.; Choi, D. G.; Tom, G.; Le, M. L.; Li, K.; Sankar, R.; Vuong, H.-T.; Wakidi, H.; Yi, A.; Zhu, Z.; Schopp, N.; Peng, A.; Luginbuhl, B. R.; Nguyen, T.-Q.; Aspuru-Guzik, A. Beyond molecular structure: Critically assessing machine learning for designing organic photovoltaic materials and devices. *J. Mater. Chem. A* **2024**, *12*, 14540−14558.
(72) An, N. G.; Kim, J. Y.; Vak, D. Machine learning-assisted development of organic photovoltaics via high-throughput in situ formulation. *Energy Environ. Sci.* **2021**, *14*, 3438−3446.
(73) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems 30*, 2017.
(74) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3*, 93.
(75) Malhotra, P.; Biswas, S.; Sharma, G. D. Directed message passing neural network for predicting power conversion efficiency in organic solar cells. *ACS Appl. Mater. Interfaces* **2023**, *15*, 37741−37747.
(76) Chen, L.-Q.; Zhang, C.-R.; Sang, C.-C.; Liu, X.-M.; Gong, J.-J.; Zhang, M.-L.; Chen, H.-S. High-Throughput Molecular Design of Donors and Non-Fullerene Acceptors for Organic Solar Cells Based on Convolutional Neural Networks. *J. Chem. Inf. Model.* **2025**, *65*, 10107−10123.
(77) Peng, S.-P.; Zhao, Y. Convolutional neural networks for the design and analysis of non-fullerene acceptors. *J. Chem. Inf. Model.* **2019**, *59*, 4993−5001.
(78) Zhang, C.-R.; Lv, L.-F.; Li, M.; Liu, X.-M.; Gong, J.-J.; Liu, Z.-J.; Wu, Y.-Z.; Chen, H.-S. High throughput molecular design of electron donors and non-fullerene acceptors using machine learning combined with substructure importance. *J. Mater. Chem. C* **2025**, *13*, 14864.
(79) Das, B.; Patrikar, K.; Keny, A. S.; Mondal, A. From fragments to function: data-driven design of high-performance non-fullerene acceptors for organic photovoltaics. *Mol. Syst. Des. Eng.* **2026**, *11*, 269−291.
(80) Zhu, Z.; Zhu, C.; Tu, Y.; Shao, T.; Wang, Y.; Liu, W.; Liu, Y.; Zang, Y.; Wei, Q.; Yan, W. Machine-learning-assisted exploration of new non-fullerene acceptors for high-efficiency organic solar cells. *Cell Rep. Phys. Sci.* **2024**, *5*, 102316.
(81) Abdelghafar, S.; Alshater, H.; Abouelmagd, L. M.; Darwish, A.; Hassanien, A. E. Organic photovoltaic prediction model based on Bayesian optimization and explainable AI. *Sci. Rep.* **2025**, *15*, 32559.
(82) Lv, L.-F.; Zhang, C.-R.; Sang, C.-C.; Liu, X.-M.; Zhang, M.-L.; Gong, J.-J.; Chen, Y.-H.; Chen, H.-S. Integrating deep learning and symbolic regression for molecular design and virtual screening of organic solar cells. *npj Comput. Mater.* **2025**, *12*, 31.
(83) Lee, M.-H. Interpretable machine learning model for the highly accurate prediction of efficiency of ternary organic solar cells based on nonfullerene acceptor using effective molecular descriptors. *Sol. RRL* **2023**, *7*, 2300307.
(84) Siddiqui, H.; Usmani, T. Interpretable AI and Machine Learning Classification for Identifying High-Efficiency Donor−Acceptor Pairs in Organic Solar Cells. *ACS Omega* **2024**, *9*, 34445−34455.
(85) Li, Q.; Wang, L.-M.; Liu, S.; Zhan, X.; Zhu, T.; Cao, Z.; Lai, H.; Zhao, J.; Cai, Y.; Xie, W.; Huang, F. Impact of donor−acceptor interaction and solvent additive on the vertical composition distribution of bulk heterojunction polymer solar cells. *ACS Appl. Mater. Interfaces* **2019**, *11*, 45979−45990.
(86) Lin, Y. L.; Fusella, M. A.; Rand, B. P. The impact of local morphology on organic donor/acceptor charge transfer states. *Adv. Energy Mater.* **2018**, *8*, 1702816.
(87) Wang, H.; Cao, J.; Yu, J.; Zhang, Z.; Geng, R.; Yang, L.; Tang, W. Molecular engineering of central fused-ring cores of non-fullerene acceptors for high-efficiency organic solar cells. *J. Mater. Chem. A* **2019**, *7*, 4313−4333.
(88) Katubi, K. M.; Saqib, M.; Mubashir, T.; Tahir, M. H.; Halawa, M. I.; Akbar, A.; Basha, B.; Sulaman, M.; Alrowaili, Z.; Al-Buriahi, M. Predicting the multiple parameters of organic acceptors through machine learning using RDkit descriptors: an easy and fast pipeline. *Int. J. Quantum Chem.* **2023**, *123*, No. e27230.
(89) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet−a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(90) Wang, H.; Feng, J.; Dong, Z.; Jin, L.; Li, M.; Yuan, J.; Li, Y. Efficient screening framework for organic solar cells with deep learning and ensemble learning. *npj Comput. Mater.* **2023**, *9*, 200.
(91) Zhang, W.; Zou, Y.; Wang, X.; Chen, J.; Xu, D. Deep learning accelerated high-throughput screening of organic solar cells. *J. Mater. Chem. C* **2025**, *13*, 5295−5306.
(92) Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. *International Conference on Artificial Neural Networks*, 2018; pp 270−279.
(93) Hussain, M.; Bird, J. J.; Faria, D. R. A study on CNN transfer learning for image classification; *Workshop on Computational Intelligence*: UK, 2018; pp 191−202.
(94) Moore, G. J.; Bardagot, O.; Banerji, N. 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.
(95) Muller, J. S.; Comi, M.; Eisner, F.; Azzouzi, M.; Herrera Ruiz, D.; Yan, J.; Attar, S. S.; Al-Hashimi, M.; Nelson, J. Charge-transfer state dissociation efficiency can limit free charge generation in low-offset organic solar cells. *ACS Energy Lett.* **2023**, *8*, 3387−3397.
(96) Cnops, K.; Zango, G.; Genoe, J.; Heremans, P.; Martinez-Diaz, M. V.; Torres, T.; Cheyns, D. Energy level tuning of non-fullerene acceptors in organic solar cells. *J. Am. Chem. Soc.* **2015**, *137*, 8991−8997.
(97) Bartesaghi, D.; Pérez, I. D. C.; Kniepert, J.; Roland, S.; Turbiez, M.; Neher, D.; Koster, L. J. A. Competition between recombination and extraction of free charges determines the fill factor of organic solar cells. *Nat. Commun.* **2015**, *6*, 7083.
(98) Zhang, G.; Chen, X.-K.; Xiao, J.; Chow, P. C.; Ren, M.; Kupgan, G.; Jiao, X.; Chan, C. C.; Du, X.; Xia, R.; Chen, Z.; Yuan, J.; Zhang, Y.; Zhang, S.; Liu, Y.; Zou, Y.; Yan, H.; Wong, K. S.; Coropceanu, V.; Li, N.; Brabec, C. J.; Bredas, J.-L.; Yip, H.-L.; Cao, Y. Delocalization of exciton and electron wavefunction in non-fullerene acceptor molecules enables efficient organic solar cells. *Nat. Commun.* **2020**, *11*, 3943.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

(99) Sahu, H.; Mahmood, A.; Shafique, L. B.; Ramprasad, R. From Corpus to Innovation: Advancing Organic Solar Cell Design with Large Language Models. *npj Comput. Mater.* **2025**, *12*, 27.
(100) Zhang, C.-R.; Li, M.; Zhao, M.; Gong, J.-J.; Liu, X.-M.; Chen, Y.-H.; Liu, Z.-J.; Wu, Y.-Z.; Chen, H.-S. Machine learning study on organic solar cells and virtual screening of designed non-fullerene acceptors. *J. Appl. Phys.* **2023**, *134*, 153104.
(101) Malhotra, P.; Verduzco, J. C.; Biswas, S.; Sharma, G. D. Active discovery of donor: acceptor combinations for efficient organic solar cells. *ACS Appl. Mater. Interfaces* **2022**, *14*, 54895−54906.
(102) Segal, N.; Netanyahu, A.; Greenman, K. P.; Agrawal, P.; Gómez-Bombarelli, R. Known unknowns: Out-of-distribution property prediction in materials and molecules. *npj Comput. Mater.* **2025**, *11*, 345.
(103) Antoniuk, E. R.; Zaman, S.; Ben-Nun, T.; Li, P.; Diffenderfer, J.; Sahin, B.; Smolenski, O.; Hsu, T.; Hiszpanski, A. M.; Chiu, K.; Kailkhura, B.; Van Essen, B. Boom: benchmarking out-of-distribution molecular property predictions of machine learning models. *arXiv* **2025**, arXiv:2505.01912.
(104) Shafian, S.; Mohd Salehin, F. N.; Lee, S.; Ismail, A.; Mohamed Shuhidan, S.; Xie, L.; Kim, K. Development of organic semiconductor materials for organic solar cells via the integration of computational quantum chemistry and AI-powered machine learning. *ACS Appl. Energy Mater.* **2025**, *8*, 699−722.
(105) Greenstein, B. L.; Hutchison, G. R. Screening efficient tandem organic solar cells with machine learning and genetic algorithms. *J. Phys. Chem. C* **2023**, *127*, 6179−6191.
(106) Khanam, L.; Srivastava, S. B.; Do, T. T.; Sonar, P.; Singh, S. P. Non-fullerene acceptor-based nanomorphology enhancement for efficient ternary organic solar cells. *Phys. Status Solidi A* **2022**, *219*, 2200143.
(107) Yu, T.; Ma, D. Highly efficient nonfullerene organic solar cells: Morphology control and characterizations. *Sol. RRL* **2024**, *8*, 2300751.
(108) An, K.; Zhong, W.; Peng, F.; Deng, W.; Shang, Y.; Quan, H.; Qiu, H.; Wang, C.; Liu, F.; Wu, H.; Li, N.; Huang, F.; Ying, L. Mastering morphology of non-fullerene acceptors towards long-term stable organic solar cells. *Nat. Commun.* **2023**, *14*, 2688.
(109) Carrillo-Sendejas, J. C.; Maldonado, J.-L. Progress in organic solar cells: Materials, challenges, and novel strategies for niche applications. *APL Energy* **2025**, *3*, 021501.
(110) Karki, A.; Vollbrecht, J.; Gillett, A. J.; Xiao, S. S.; Yang, Y.; Peng, Z.; Schopp, N.; Dixon, A. L.; Yoon, S.; Schrock, M.; Ade, H.; Reddy, G. N. M.; Friend, R. H.; Nguyen, T.-Q. The role of bulk and interfacial morphology in charge generation, recombination, and extraction in non-fullerene acceptor organic solar cells. *Energy Environ. Sci.* **2020**, *13*, 3679−3692.
(111) Zhu, L.; Zhang, M.; Zhong, W.; Leng, S.; Zhou, G.; Zou, Y.; Su, X.; Ding, H.; Gu, P.; Liu, F.; Zhang, Y. Progress and prospects of the morphology of non-fullerene acceptor based high-efficiency organic solar cells. *Energy Environ. Sci.* **2021**, *14*, 4341−4357.
(112) Chen, Q.; Wang, W.; Liu, X.; Iqbal, S.; Wang, Z. Advancements in morphology controllable ternary organic solar cells for active layers. *Renewable Sustainable Energy Rev.* **2025**, *216*, 115673.
(113) Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206−215.
(114) Zhong, X.; Gallagher, B.; Liu, S.; Kailkhura, B.; Hiszpanski, A.; Han, T. Y.-J. Explainable machine learning in materials science. *npj Comput. Mater.* **2022**, *8*, 204.
(115) Cao, R.; Zhang, C.-R.; Liu, X.-M.; Gong, J.-J.; Zhang, M.-L.; Liu, Z.-J.; Wu, Y.-Z.; Chen, H.-S. Molecular design of organic photovoltaic donors and non-fullerene acceptors: a combined machine learning and genetic algorithm approach. *J. Mater. Chem. C* **2025**, *13*, 12150−12168.
(116) Kwak, H. S.; An, Y.; Giesen, D. J.; Hughes, T. F.; Brown, C. T.; Leswing, K.; Abroshan, H.; Halls, M. D. Design of organic electronic materials with a goal-directed generative model powered by deep neural networks and high-throughput molecular simulations. *Front. Chem.* **2022**, *9*, 800370.
(117) Ogbaje, M.; Bhat, V.; Risko, C. Advances in the Design and Discovery of Organic Semiconductors Aided by Machine Learning. *Annu. Rev. Mater. Res.* **2025**, *55*, 285−306.
(118) Lv, L.-F.; Zhang, C.-R.; Cao, R.; Liu, X.-M.; Zhang, M.-L.; Gong, J.-J.; Liu, Z.-J.; Wu, Y.-Z.; Chen, H.-S. Design and virtual screening of donor and non-fullerene acceptor for organic solar cells using long short-term memory model. *J. Mater. Chem. A* **2024**, *12*, 23859−23871.
(119) Flam-Shepherd, D.; Zhigalin, A.; Aspuru-Guzik, A. Scalable fragment-based 3d molecular design with reinforcement learning. *arXiv* **2022**, arXiv:2202.00658.
(120) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv* **2017**, arXiv:1705.10843.
(121) Jiang, J.; Ke, L.; Chen, L.; Dou, B.; Zhu, Y.; Liu, J.; Zhang, B.; Zhou, T.; Wei, G.-W. Transformer technology in molecular science. *WIREs Comput. Mol. Sci.* **2024**, *14*, No. e1725.
(122) Sultan, A.; Sieg, J.; Mathea, M.; Volkamer, A. Transformers for molecular property prediction: Lessons learned from the past five years. *J. Chem. Inf. Model.* **2024**, *64*, 6259−6280.
(123) Kwak, B.; Park, J.; Kang, T.; Jo, J.; Lee, B.; Yoon, S. GeoT: a geometry-aware transformer for reliable molecular property prediction and chemically interpretable representation learning. *ACS Omega* **2023**, *8*, 39759−39769.
(124) Qiu, J.; Lam, H. H.; Hu, X.; Li, W.; Fu, S.; Zeng, F.; Zhang, H.; Wang, X. Accelerating High-Efficiency Organic Photovoltaic Discovery via Pretrained Graph Neural Networks and Generative Reinforcement Learning. *arXiv* **2025**, arXiv:2503.23766.
(125) Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. d. P.; Aspuru-Guzik, A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **2021**, *12*, 7079−7090.
(126) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
(127) Tan, J. D.; Ramalingam, B.; Chellappan, V.; Gupta, N. K.; Dillard, L.; Khan, S. A.; Galvin, C.; Hippalgaonkar, K. Generative design and experimental validation of non-fullerene acceptors for photovoltaics. *ACS Energy Lett.* **2024**, *9*, 5240−5250.
(128) Xu, P.; Ma, Y.; Lu, W.; Li, M.; Zhao, W.; Dai, Z. Multi-objective optimization in machine learning assisted materials design and discovery. *J. Mater. Inform.* **2025**, *5*, 174107.
(129) Greenstein, B. L.; Hiener, D. C.; Hutchison, G. R. Computational evolution of high-performing unfused non-fullerene acceptors for organic solar cells. *J. Chem. Phys.* **2022**, *156*, 174107.
(130) Ardayfio, C. Computational design of organic solar cell active layer through genetic algorithm. *arXiv* **2019**, arXiv:1910.12401.
(131) On, Y.; Kim, S.; Kim, S. Multi-Objective optimization for design of an Agrophotovoltaic system under Non-Dominated sorting Genetic algorithm II. *Comput. Electron. Agric.* **2024**, *224*, 109237.
(132) Markina, A.; Lin, K.-H.; Liu, W.; Poelking, C.; Firdaus, Y.; Villalva, D. R.; Khan, J. I.; Paleti, S. H.; Harrison, G. T.; Gorenflot, J.; Zhang, W.; De Wolf, S.; McCulloch, I.; Anthopoulos, T. D.; Baran, D.; Laquai, F.; Andrienko, D. Chemical Design Rules for Non-Fullerene Acceptors in Organic Solar Cells. *Adv. Energy Mater.* **2021**, *11*, 2102363.
(133) Khatua, R.; Das, B.; Mondal, A. Rational design of non-fullerene acceptors via side-chain and terminal group engineering: a computational study. *Phys. Chem. Chem. Phys.* **2023**, *25*, 7994−8004.
(134) Li, J.; Tao, Y.; Chen, S.; Li, H.; Chen, P.; Wei, M.-z.; Wang, H.; Li, K.; Mazzeo, M.; Duan, Y. A flexible plasma-treated silver-nanowire electrode for organic light-emitting devices. *Sci. Rep.* **2017**, *7*, 16468.
(135) Wadsworth, A.; Moser, M.; Marks, A.; Little, M. S.; Gasparini, N.; Brabec, C. J.; Baran, D.; McCulloch, I. 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.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX

(136) Ye, L.; Weng, K.; Xu, J.; Du, X.; Chandrabose, S.; Chen, K.; Zhou, J.; Han, G.; Tan, S.; Xie, Z.; Yi, Y.; Li, N.; Liu, F.; Hodgkiss, J. M.; Brabec, C. J.; Sun, Y. Unraveling the influence of non-fullerene acceptor molecular packing on photovoltaic performance of organic solar cells. *Nat. Commun.* **2020**, *11*, 6005.
(137) Steinmann, S. N.; Corminboeuf, C. Exploring the limits of density functional approximations for interaction energies of molecular precursors to organic electronics. *J. Chem. Theory Comput.* **2012**, *8*, 4305−4316.
(138) Hermann, J.; Tkatchenko, A. Density functional model for van der Waals interactions: Unifying many-body atomic approaches with nonlocal functionals. *Phys. Rev. Lett.* **2020**, *124*, 146401.
(139) Wildman, J.; Repiscak, P.; Paterson, M. J.; Galbraith, I. General force-field parametrization scheme for molecular dynamics simulations of conjugated materials in solution. *J. Chem. Theory Comput.* **2016**, *12*, 3813−3824.
(140) Kupgan, G.; Chen, X.-K.; Brédas, J.-L. Molecular packing of non-fullerene acceptors for organic solar cells: Distinctive local morphology in Y6 vs. ITIC derivatives. *Mater. Today Adv.* **2021**, *11*, 100154.
(141) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **2021**, *7*, 1356−1367.
(142) Gruich, C. J.; Madhavan, V.; Wang, Y.; Goldsmith, B. R. Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. *Mach. Learn.: Sci. Technol.* **2023**, *4*, 025019.
(143) Lu, Z.; An, C.; Liu, X.; Mei, Z.; Xie, X.; Li, K.; Wu, Y.; Liao, Q.; Fu, H. Implementing High-Throughput Screening of Organic Solar Cells using Transfer Learning Based on Fine-Tuning Neural Network Strategy. *Adv. Opt. Mater.* **2025**, *13*, 2402405.
(144) Kaya, M.; Hajimirza, S. Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies. *Sci. Rep.* **2019**, *9*, 5034.
(145) Yoshida, N.; Iwabuchi, Y.; Igarashi, Y.; Iwasaki, Y. Networking autonomous material exploration systems through transfer learning. *npj Comput. Mater.* **2025**, *11*, 362.
(146) Lam, H. H.; Qiu, J.; Hu, X.; Li, W.; Zeng, F.; Fu, S.; Zhang, H.; Wang, X. CycleChemist: A Dual-Pronged Machine Learning Framework for Organic Photovoltaic Discovery. *arXiv* **2025**, arXiv:2511.19500.
(147) Kunkel, C.; Margraf, J. T.; Chen, K.; Oberhofer, H.; Reuter, K. Active discovery of organic semiconductors. *Nat. Commun.* **2021**, *12*, 2422.
(148) Hußner, M.; Pacalaj, R. A.; Müller-Dieckert, G. O.; Liu, C.; Zhou, Z.; Majeed, N.; Greedy, S.; Ramirez, I.; Li, N.; Hosseini, S. M.; Uhrich, C.; Brabec, C. J.; Durrant, J. R.; Deibel, C.; MacKenzie, R. C. I. Machine learning for ultra high throughput screening of organic solar cells: solving the needle in the haystack problem. *Adv. Energy Mater.* **2024**, *14*, 2303000.
(149) Shetty, P.; Adeboye, A.; Gupta, S.; Zhang, C.; Ramprasad, R. Accelerating materials discovery for polymer solar cells: Data-driven insights enabled by natural language processing. *Chem. Mater.* **2024**, *36*, 7676−7689.
(150) Han, G.; Guo, Y.; Song, X.; Wang, Y.; Yi, Y. Terminal $\pi-\pi$ stacking determines three-dimensional molecular packing and isotropic charge transport in an A−π−A electron acceptor for non-fullerene organic solar cells. *J. Mater. Chem. C* **2017**, *5*, 4852−4857.
(151) Aldrich, T. J.; Matta, M.; Zhu, W.; Swick, S. M.; Stern, C. L.; Schatz, G. C.; Facchetti, A.; Melkonyan, F. S.; Marks, T. J. Fluorination effects on indacenodithienothiophene acceptor packing and electronic structure, end-group redistribution, and solar cell photovoltaic response. *J. Am. Chem. Soc.* **2019**, *141*, 3274−3287.
(152) Mi, D.; Kim, J.-H.; Kim, H. U.; Xu, F.; Hwang, D.-H. Fullerene derivatives as electron acceptors for organic photovoltaic cells. *J. Nanosci. Nanotechnol.* **2014**, *14*, 1064−1084.
(153) Nielsen, C. B.; Holliday, S.; Chen, H.-Y.; Cryer, S. J.; McCulloch, I. Non-fullerene electron acceptors for use in organic solar cells. *Acc. Chem. Res.* **2015**, *48*, 2803−2812.
(154) Bai, H.; Wu, Y.; Wang, Y.; Wu, Y.; Li, R.; Cheng, P.; Zhang, M.; Wang, J.; Ma, W.; Zhan, X. Nonfullerene acceptors based on extended fused rings flanked with benzothiadiazolylmethylenemalononitrile for polymer solar cells. *J. Mater. Chem. A* **2015**, *3*, 20758−20766.

https://doi.org/10.1021/acsomega.6c01194
ACS Omega XXXX, XXX, XXX−XXX