Electronic Supplementary Material

# Developing an Intelligent Data-Driven Framework for Organic Photovoltaic Research

*Yu Cui*, [a] *Wei Ma,* [a] *and Han Yan** [a]

[a]State Key Laboratory for Mechanical Behavior of Materials; Xi'an Jiaotong University; Xi'an 710049, P. R. China. E-mail: mseyanhan@mail.xjtu.edu.cn

## Experimental:

### Large language model and evaluation:

*Large language model:* This study employs three state-of-the-art open-source large language models (LLMs) to process textual content, tables and figures for information extraction: Qwen2.5-72B, Qwen2.5-VL-72B, and DeepSeek-V3. The Qwen2.5-72B and Qwen2.5-VL-72B (2024) models were developed by Tongyi Lab, part of Alibaba Group, as members of the Qwen series; DeepSeek-V3 was developed by DeepSeek. All three models are built upon the Transformer architecture and operate in an autoregressive manner for text generation. Specifically, Qwen2.5-72B and DeepSeek-V3 are unimodal language models, while Qwen2.5-VL-72B is a multimodal extension capable of joint understanding of figures and text. For experimental evaluation, we deployed these open-source models on cloud-based servers equipped with high-performance GPUs. Efficient inference was achieved using the Hugging Face Transformers library, and model integration with downstream tasks was implemented via custom API interfaces. Code is uploaded to GitHub (https://github.com/limitedcommunication/Data-extraction)

*Evaluation of Text Mining:* To evaluate the performance of LLMs in extracting parameters from OPV literature, precision, recall, and F1 score were primarily employed as the performance metrics, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

In the text mining task for OPV design parameters, each extracted parameter was classified into one of the following three labels: TP (accurately extracted parameter), FP (incorrectly extracted parameter or irrelevant information extracted), or FN (parameter not successfully extracted). The precision metric measured the accuracy of the method in extracting OPV parameters, the recall metric assessed the completeness of the method in extracting OPV parameters, and the F1 score, which was calculated

based on precision and recall, served as an overall representation of the method's performance.

**Calculation platform construction and molecular parser:**

*Molecular Parser:* The molecular parser utilizes the DECIMER (version 2.7.1) open-source platform, which leverages cutting-edge advances in deep learning, computer vision, and natural language processing to automatically segment, classify, and translate chemical structure depictions from printed literature. This platform enables seamless conversion of molecular structure images in PDF documents into SMILES strings.

In our study, DECIMER was employed exclusively to interpret chemical structure diagrams of Y-series NFAs and convert them into machine-readable SMILES representations. Applied to the 125 molecules in our dataset, DECIMER successfully generated chemically valid and accurate SMILES strings for 110 molecules (88%). The remaining 15 molecules (12%) featured low-resolution or ambiguous structural diagrams, which led to incorrect SMILES outputs—most commonly misassigned alkyl side chains. For these cases, expert chemists manually reviewed the original publications and corrected the SMILES strings to ensure chemical fidelity. This resulted in a manual curation rate of 12%, which reflects the current limitations of automated chemical image recognition in real-world scientific literature but also underscores the necessity of domain expertise in high-stakes data curation.

*High-Throughput Quantum Chemistry Computation Platform (HQCCP):*

(1) A molecular processing pipeline was constructed using the RDKit (version 2025.3.3) cheminformatics toolkit, supporting automated conversion and standardization of multiple molecular input formats including SMILES, SDF, and XYZ.

(2) Implementation of tiered computational strategy:

(a) Primary screening: GFN2-xTBsemi-empirical method on clusters using xTB.[1,2]

(b) Secondary refinement: Geometry optimization using ORCA 5.0 with B3LYP

functional and def2-SVP basis set, employing TIGHTSCF integration accuracy keywords.[3-7]

(c) Tertiary high-accuracy calculation: Single-point energy correction for selected candidates at the same level of theory, accelerated via RIJCOSX approximation.[8]

(3) Computational task management system:

Automated scheduling framework developed in Python 3.9. Independent monitoring threads for each job to track convergence status and errors in .out files. Automatic adjustment of convergence thresholds or damping algorithms for SCF failures. Three-level retry mechanism with automatic logging of problematic molecules after maximum attempts

(4) Data acquisition and processing workflow:

Molecular descriptors calculated using Multiwfn 3.8(dev).[9] Multithreaded result parser based on cclib extracts 50 molecule descriptors. Results stored in standardized CSV format.

This platform achieves standardized computational workflows, intelligent task management, and systematic data analysis through the above technical approaches. Code is uploaded to GitHub (https://github.com/limitedcommunication/HQCCP)

**Machine learning tools:**

*XGBoost (eXtreme Gradient Boosting)* is a powerful ensemble machine learning algorithm widely used for regression and classification tasks due to its efficiency, scalability, and performance. It is based on gradient boosting frameworks that sequentially build decision trees, where each subsequent tree aims to correct the errors of the previous ones. In this study, we employed the XGBoost algorithm (XGBoost Python package, version 2.0.0) to predict OPV device performance from molecular descriptors. XGBoost introduces several improvements over traditional gradient boosting, including L1/L2 regularization to prevent overfitting, built-in handling of missing values, and support for parallel processing to accelerate training.

Hyperparameter ranges and types used in our research:

learning_rate: [0.01, 0.05, 0.1, 0.2]

max_depth: [3, 4, 5, 6, 7]

n_estimators: [100, 200, 300, 500]

subsample: [0.8, 0.9, 1.0]

colsample_bytree: [0.8, 0.9, 1.0]

gamma: [0, 0.1, 0.2, 0.5]

The best-performing combination (selected based on mean validation $r^2$ across the 5 folds) was: learning_rate=0.05, max_depth=5, n_estimators=300, subsample=0.9, colsample_bytree=0.9, gamma=0.1.

*Valiadation:* To validate the accuracy and generalisability of the machine learning model, we computed 27 molecular descriptors for L8-BO-X and input them into the pre-trained model, ultimately yielding a predicted PCE of 17.55%. The codes and documents were available on GitHub (https://github.com/limitedcommunication/L8-BO-X_validation).

*SHapley Additive exPlanations (SHAP)* is a unified and interpretable game theory-based framework for explaining the output of any machine learning model. It assigns each feature an importance value for a particular prediction, reflecting its contribution to the model's decision. SHAP values are derived from Shapley values in cooperative game theory, ensuring fair attribution of feature contributions by considering all possible feature combinations. In this study, SHAP (version 0.42.1) was applied post hoc to the trained XGBoost model to quantify the impact of individual molecular descriptors on predicted OPV device performance. This approach not only enabled global interpretation of feature importance across the entire dataset but also facilitated local explanation of individual predictions, revealing how specific molecular characteristics influence device efficiency under different structural contexts. By integrating SHAP with ECFP-based fragment analysis, we further decoded structure–property relationships at the submolecular level, supporting rational molecular design strategies. The codes were available on GitHub (https://github.com/limitedcommunication/OPV_analyzer).

*Extended Connectivity Fingerprints (ECFP)* are a class of circular topological molecular fingerprints widely used in cheminformatics and machine learning

applications for drug discovery and materials science. ECFP encodes the local atomic environment around each atom in a molecule by iteratively considering increasing layers of neighboring atoms up to a specified diameter. Calculation steps:

(1) Initialize Atom Identifiers.

Generate initial identifiers for each atom in the molecule using a set of basic atomic properties, typically including atom type, hybridization state, formal charge, etc.

(2) Iterative Atomic Environment Expansion.

Starting from the target atom, iteratively expand the atomic environment outward based on a defined radius, collecting information about surrounding atoms and bonds (Figure S10).

1. Iteration 0 (Initial Atom Identifier): The identifier represents only Atom 1 and its directly bonded atoms (smallest circle).
2. Iteration 1: The identifier now includes information about Atom 1's immediate neighboring atoms (intermediate circle).
3. Iteration 2: The substructure further expands, covering most of the terminal group (largest circle).

(3) Hashing (Fingerprint Encoding).

To compress these complex structural identifiers into a fixed-length fingerprint, apply a hashing function to each substructure identifier. Each hash value corresponds to a specific position (bit or feature) in the fingerprint vector.

(4) Generate Final Fingerprint.

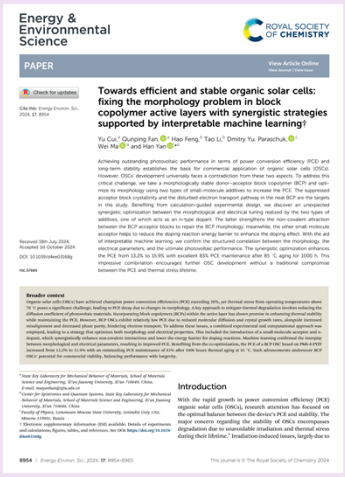
Based on the hashing results, set the value of each bit/feature to 0 (absent) or 1 (present) to indicate the occurrence of a given substructure.

In this study, ECFP fingerprints were generated using the RDKit package (version 2022.09.5), we set radius to 2 and nBits to 1024 and employed alongside quantum chemistry-derived molecular descriptors to enhance model interpretability and uncover fragment-level insights into molecular design principles influencing OPV performance.

# Figures:

**Prompt:**
Extract the article information from the image, including title, authors,abstract, journal and publication year. Output in JSON format with the following specifications:
1. Title: The complete article title, appearing as larger bold text at the top of the image.
2. Authors: Extract authors separated by comma/semicolon below the title.
3. Abstract: Locate paragraphs beginning with 'Abstract', usually below the author, retaining line breaks and complete sentences.
4. Journal: The full journal name where the article was published, located in the upper right/lesft section usually.
5. Publication Date: The year of publication.

Output strictly in JSON format without any additional comments or punctuation:
{
"title": "Article title",
"authors": "Author names",
"abstract": "Extracted abstract",
"journal": "Journal name",
"year": "Publication year"
}


Energy & Environmental Science

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online

Check for updates

Towards efficient and stable organic solar cells: fixing the morphology problem in block copolymer active layers with synergistic strategies supported by interpretable machine learning†

Yu Cui,‡ Qunping Fan,‡ Hao Feng,‡ Tao Li, Dmitry Yu. Paraschuk, Wei Ma, Han Yan

Broader context

Introduction

Output:

{
"title": "Towards efficient and stable organic solar cells: fixing the morphology problem in block copolymer active layers with synergistic strategies supported by interpretable machine learning",
"authors": "Yu Cui, Qunping Fan, Hao Feng, Tao Li, Dmitry Yu. Paraschuk, Wei Ma, Han Yan",
"abstract": "Achieving outstanding photovoltaic performance in terms of power conversion efficiency (PCE) and long-term stability establishes the basis for commercial application of organic solar cells (OSCs). However, OSCs' development universally faces a contradiction from these two aspects. To address this critical challenge, we take a morphologically stable donor–acceptor block copolymer (BCP) and optimize its morphology using two types of small-molecule additives to increase the PCE. The suppressed acceptor block crystallinity and the disturbed electron transport pathway in the neat BCP are the targets in this study. Benefiting from calculation-guided experimental design, we discover an unexpected synergistic optimization between the morphological and electrical tuning realized by the two types of additives, one of which acts as an n-type dopant. The latter strengthens the non-covalent attraction between the BCP acceptor blocks to repair the BCP morphology; meanwhile, the other small-molecule acceptor helps to reduce the doping reaction energy barrier to enhance the doping effect. With the aid of interpretable machine learning, we confirm the structured correlation between the morphology, the electrical parameters, and the ultimate photovoltaic performance. The synergistic optimization enhances the PCE from 13.2% to 15.9% with excellent 83% PCE maintenance after 85 °C aging for 1000 h. This impressive combination encourages further OSC development without a traditional compromise between the PCE and thermal stress lifetime.",
"journal": "Energy & Environmental Science",
"year": "2024"
}


**Fig. S1.** An example of using the open-source multimodal LLM Qwen2.5-VL-72B to extract abstract and other information from a screenshot image of scientific literature.

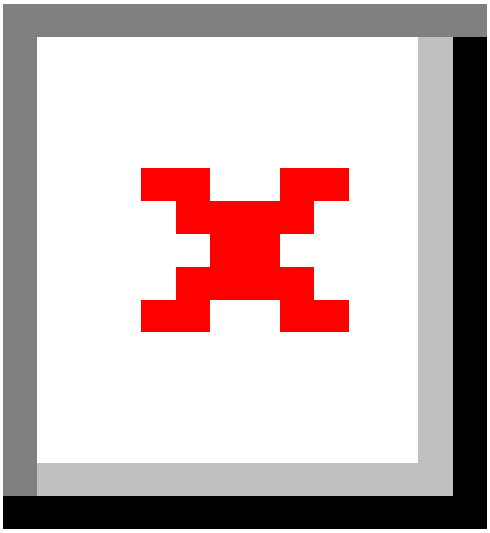


**Fig. S2.** An example of using the open-source LLM DeepSeek-V3 to classify scientific literature based on extracted abstracts.

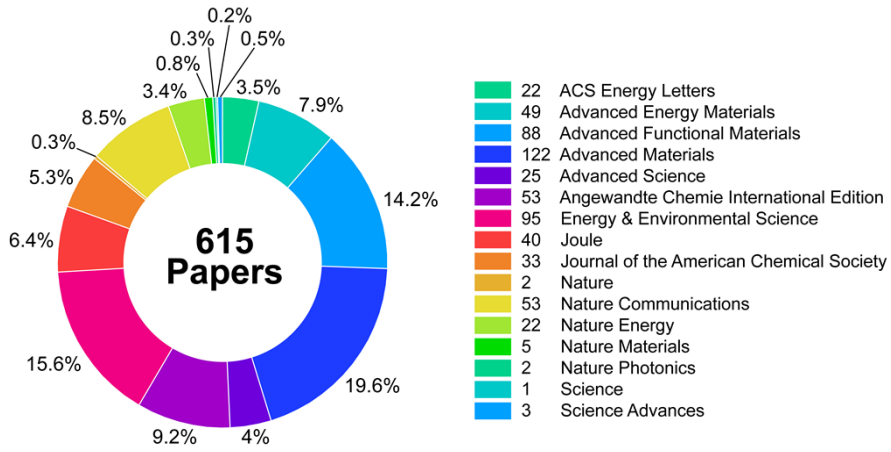



**Fig. S3.** Distribution of the academic journals of the literatures used in this study.

**Material Pairs Extraction Task Prompt:**

```
Task Objective:
Extract the corresponding material combinations for the following PV performance conditions:
{performance_conditions}
 Extraction rules
Combination type definition
- binary
  Structure: 1 donor (D1) + 1 acceptor (A1).
  Example: `D1/A1`.
- Ternary
  Structure:
  - 2 donor(D1/D2) + 1 acceptor(A1)
  - 1-donor(D1) + 2-acceptor(A1/A2)
- Multi-component
  Number of components: >3
- Single-component(single-component)
  Type:
  - Donor-acceptor copolymer (e.g. `PM6-b-PYIT`)
  - Pure donor/pure acceptor material
Property requirements
- Donor-acceptor ratio
  Mark 'null' if unspecified.
- Additives
  Includes: liquid/solid additives, organic dopants
  Mark 'not mentioned' if unspecified
- Extraction range
  Coverage: text/tables/footnotes
- Experimental condition discrimination
  Include: same material combinations, but different experimental conditions are classified as different condition_ids
- Output format:
  - Output in JSON format with material combination type, donor and acceptor material names.
Output in JSON format only, without any additional comments, symbols, or explanations.
Output format:
{
  'conditions": [
    {
      'condition_id": "perf_01",
      'materials": {
        'type": "binary/ternary/etc",
        'donors': ['donor material"],
        'acceptors': ['acceptor material"],
        'DA_ratio": "DA ratio",
        'additives': ['Additives 1']
      }
    }
  ]
}
```

**Processing Conditional Extraction Task Prompt:**

```
Task Objective
Extract the corresponding processing conditions for the following PV performance conditions:
{performance_conditions}
Extraction rules
Range of processing conditions
- Core parameters:
  - Solvent
  - Device structure
  - Processing method (spin coating/scratch coating/R2R, etc.)
  - Heat treatment conditions
  - Effective area
  - Film thickness
Extraction Logic
1. **Explicit conditions
   - Explicitly mentioned in the literature → extract specific values
   - Not explicitly mentioned → marked 'null'.
2. Relevance judgement
   - Verify device structure consistency
   - Verify that the effective area is the same
   - Check solvent usage
   - Verify film thickness parameters
3. Repeat condition processing
   Identical conditions need to be output independently to each corresponding condition
Data source
Coverage: All content in text/tables/footnotes etc.
Output in JSON format only, without adding any additional comments, symbols or explanations.
Output format:
{
  "conditions": [
    {
      "condition_id": "perf_01",
      "processing_conditions": {
        "structure": "device structure",
        "solvent": "solvent",
        "processing_method": "processing method",
        "thermal_annealing": "thermal annealing",
        "active_area": "active area",
        "film_thickness": "thickness"
      }
    }
  ]
}
```

**Fig. S4.** Prompts for driving DeepSeek-V3 to extract material pairs and processing conditions.

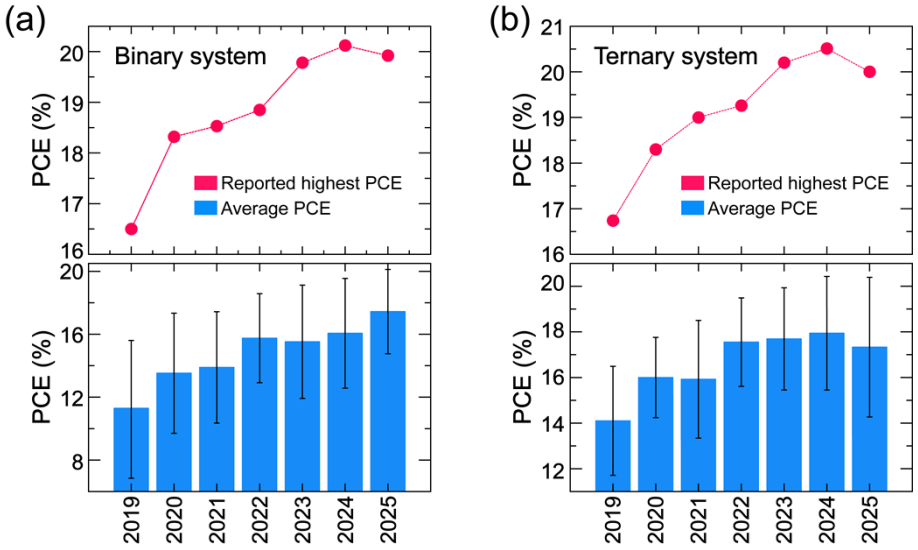



**Fig. S5.** From 2019 to 2025, (a) changes in the distribution of reported maximum and average PCE for binary systems; (b) changes in the distribution of reported maximum and average PCE for ternary systems.

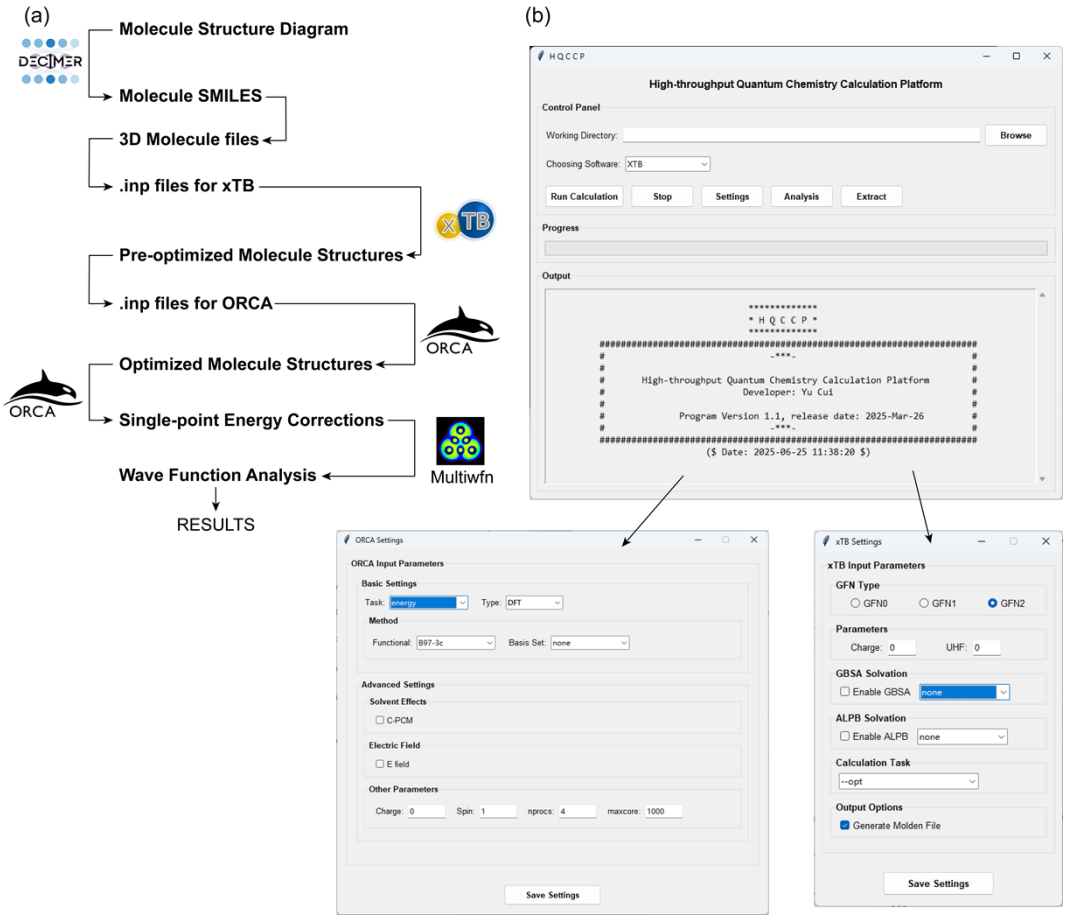



**Fig. S6.** (a) Workflow of the HQCCP. (b) Main work panel of the HQCCP with ORCA and xTB software calculation setup page.

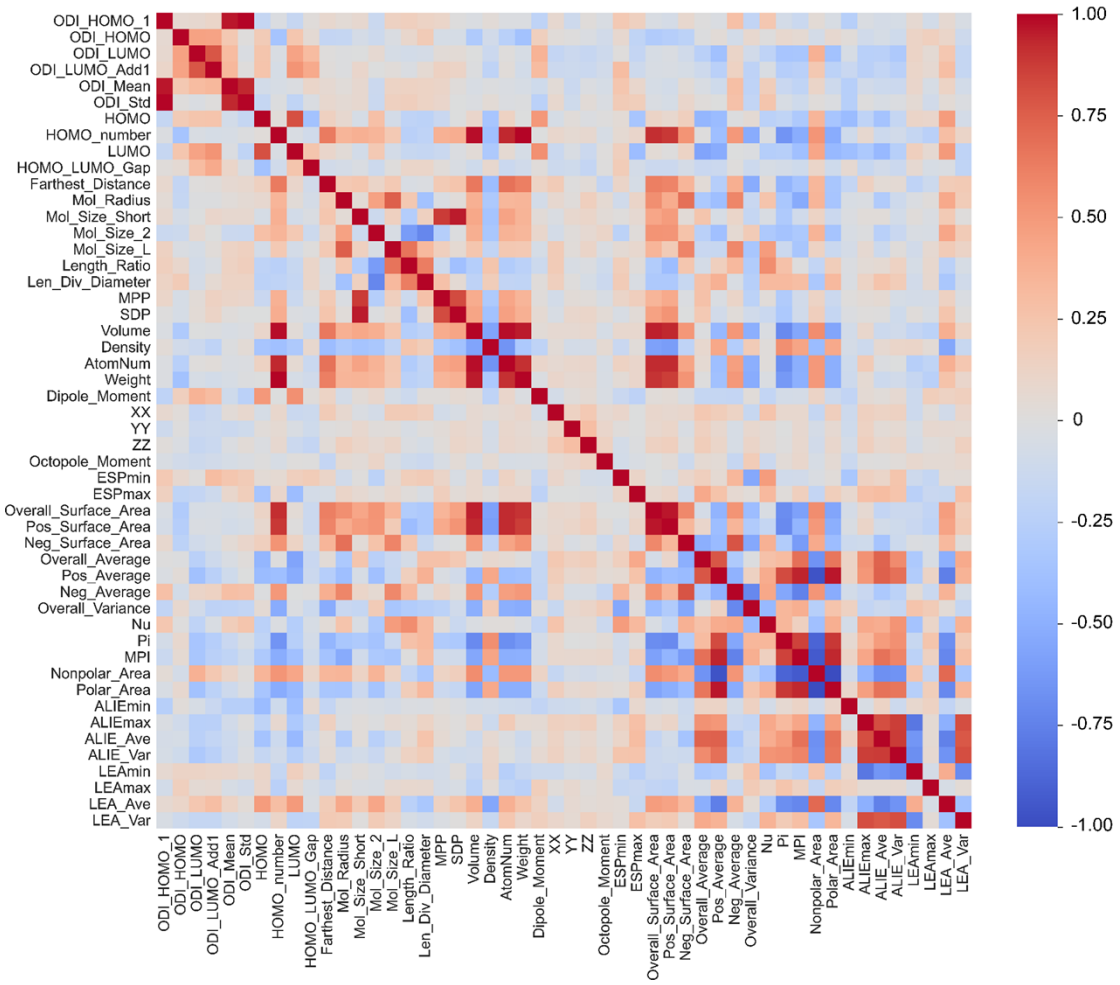



**Fig. S7.** Pearson correlation heat map of 50 molecular descriptors.

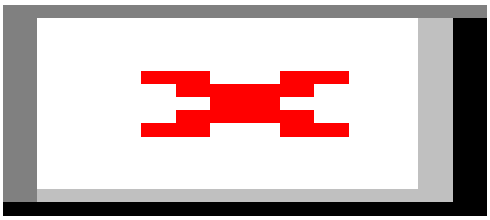


**Fig. S8.** PCE distribution of devices derived from 125 Y-series acceptors.

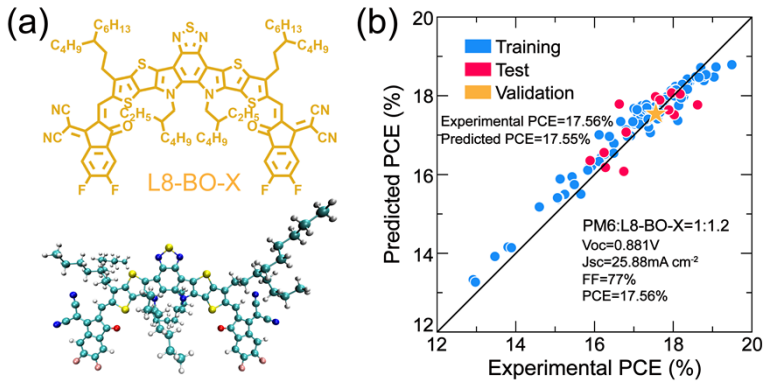



**Fig. S9.** (a) Chemical structure of L8-BO-X, including its chemical formula and DFT-optimized 3D conformation. (b) Scatter plot comparing experimental PCE versus predicted PCE for the training set (blue), test set (red), and validation set (yellow) of the XGBoost model. The diagonal line represents perfect prediction. The yellow star highlights the prediction for L8-BO-X, which achieves an experimental PCE of 17.56% and a predicted PCE of 17.55%.

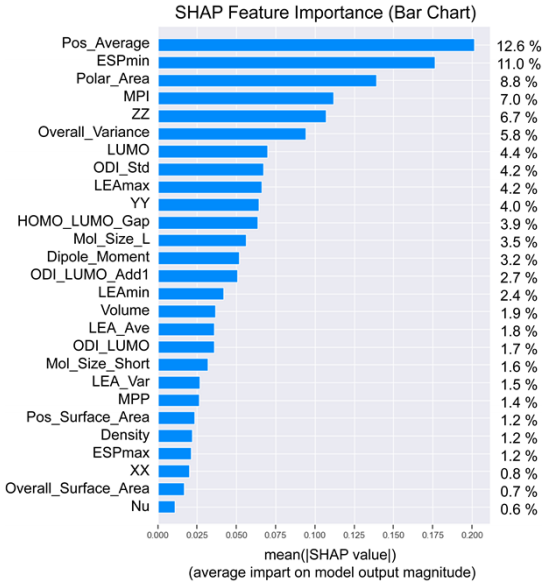



**Fig. S10.** SHAP feature importance and contribution of 27 descriptors obtained by explainable ML.

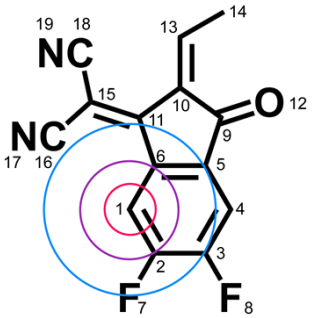



**Fig. S11.** Schematic diagram of the ECFP fingerprint perception radius. As an example, atom 1 in the terminal group of the Y-series acceptor molecule in the picture is the target atom. Where the red, purple, and blue circles represent different ranges of iterative atomic environment expansion, respectively. See calculation details in Method.

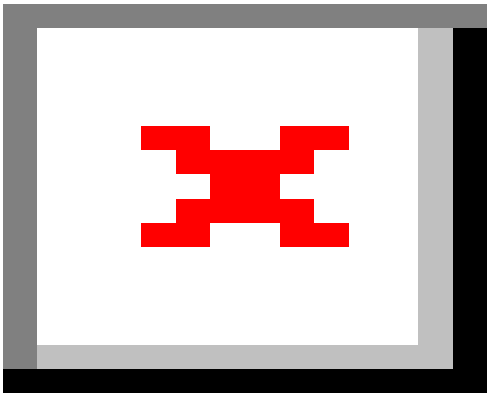


**Fig. S12.** Predicted (a) Pos_Average, (b) $ESP_{min}$, (c) Polar_Area, (d) MPI, and (e) ZZ by machine learning versus calculated results. SHAP feature importance and contribution of (f) Pos_Average, (g) ESPmin, (h) Polar_Area, (i) MPI, and (j) ZZ obtained by explainable ML. Illustration of features contributing to (k) Pos_Average, (i) ESPmin, (m) Polar_Area, (n) MPI and, (o) ZZ by SHAP values of different molecular substructures.

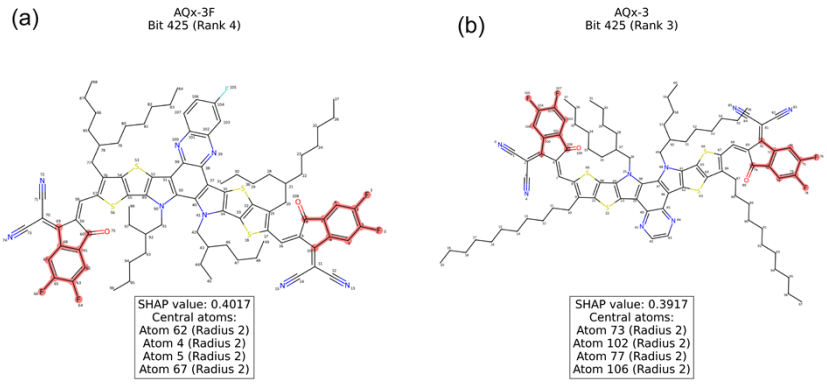



**Fig. S13.** Visualization of the atomic environment corresponding to ECFP bit 425, which exhibits a strong positive association with PCE. The red-highlighted atoms and bonds represent the substructure activated by this bit. (a) In molecule AQx-3F (Rank 4), the bit is activated by central atoms 62, 4, 5, and 67, yielding a SHAP value of +0.4017. (b) In molecule AQx-3 (Rank 3), the bit is activated by central atoms 73, 102, 77, and 106, yielding a SHAP value of +0.3917. This demonstrates that the same ECFP bit consistently identifies the ortho-fluorinated phenyl ring motif across different Y-series acceptors, confirming its role as a key structural feature for high performance.

# Tables:

**Table S1.** Comparison table of 32 parameters abbreviations and full names.

| Number | Parameter abbreviation | Full name of parameter |
|---|---|---|
| 1 | material_type | Material Type (binary or ternary etc.) |
| 2 | donors | Donor Materials |
| 3 | acceptors | Acceptor Materials |
| 4 | DA_ratio | Donor-Acceptor Ratio |
| 5 | additives | Additives |
| 6 | structure | Device Structure |
| 7 | solvent | Solvent |
| 8 | processing_method | Processing Method (spin-coating etc.) |
| 9 | thermal_annealing | Thermal Annealing (temperature and time) |
| 10 | active_area | Active Area |
| 11 | film_thickness | Active Layer Film Thickness |
| 12 | Voc_best | Best Open-Circuit Voltage |
| 13 | Voc_avg | Average Open-Circuit Voltage |
| 14 | Voc_std | Standard Deviation of Open-Circuit Voltage |
| 15 | Voc_unit | Unit of Open-Circuit Voltage |
| 16 | Jsc_best | Best Short-Circuit Current Density |
| 17 | Jsc_avg | Average Short-Circuit Current Density |
| 18 | Jsc_std | Standard Deviation of Short-Circuit Current Density |
| 19 | Jsc_unit | Unit of Short-Circuit Current Density |
| 20 | FF_best | Best Fill Factor |
| 21 | FF_avg | Average Fill Factor |
| 22 | FF_std | Standard Deviation of Fill Factor |
| 23 | FF_unit | Unit of Fill Factor |
| 24 | PCE_best | Best Power Conversion Efficiency |
| 25 | PCE_avg | Average Power Conversion Efficiency |
| 26 | PCE_std | Standard Deviation of Power Conversion Efficiency |
| 27 | PCE_unit | Unit of Power Conversion Efficiency |
| 28 | certified_status | Certified Status |
| 29 | certifying_body | Certifying Body |
| 30 | light_spectrum | Light Spectrum |
| 31 | light_intensity_value | Light Intensity Value |
| 32 | light_intensity_unit | Light Intensity Unit |

**Table S2.** Summary of 50 descriptor abbreviations and full names, and 27 were selected after screening. An asterisk (*) indicates a descriptor included in the final selected set.

| Number | Categories | Sub-Categories | Descriptor | Full name of descriptor |
|---|---|---|---|---|
| 1 | Frontier Orbitals | Frontier Orbital Energy Levels | HOMO | Highest Occupied Molecular Orbital |
| 2 | | | HOMO_number | HOMO Number |
| 3 | | | LUMO* | Lowest Unoccupied Molecular Orbital |
| 4 | | Energy Gap Metric | HOMO_LUMO_Gap* | HOMO-LUMO Gap |
| 5 | Orbital Delocalization | Frontier Orbital-Specific Delocalization Indices | ODI_HOMO_1 | Orbital Delocalization Index HOMO-1 |
| 6 | | | ODI_HOMO | Orbital Delocalization Index HOMO |
| 7 | | | ODI_LUMO* | Orbital Delocalization Index LUMO |
| 8 | | | ODI_LUMO_Add1* | Orbital Delocalization Index LUMO+1 |
| 9 | | Statistical Summaries of Orbital Delocalization | ODI_Mean | Orbital Delocalization Index Mean |
| 10 | | | ODI_Std* | Orbital Delocalization Index Standard Deviation |
| 11 | Molecular Geometry | Global Size & Spatial Extent | Farthest_Distance | Farthest Distance between Atoms |
| 12 | | | Mol_Radius | Molecular Radius |
| 13 | | | Mol_Size_Short* | Shortest Molecular Size |
| 14 | | | Mol_Size_2 | Medium Molecular Size |
| 15 | | | Mol_Size_L* | Longest Molecular Size |
| | | | MPP* | Molecular Planarity Parameter |
| | | | SDP | Span of Deviation from Plane |
| 16 | | Molecular Shape Anisotropy | Length_Ratio | Length Ratio |
| 17 | | | Len_Div_Diameter | Length Divided by Diameter |
| 20 | | Mass & Density Properties | Volume* | Molecular Volume |
| 21 | | | Density* | Molecular Density |
| 22 | | | AtomNum | Atom Numbers |
| 23 | | | Weight | Molecule Weight |
| 24 | Electrostatics & Polarity | Multipole Moments | Dipole_Moment* | Dipole Moment |
| 25 | | | XX* | Quadrupole Moment XX direction |
| 26 | | | YY* | Quadrupole Moment YY direction |
| 27 | | | ZZ* | Quadrupole Moment XX direction |
| 28 | | | Octopole_Moment | Octopole Moment |
| 29 | | Electrostatic Potential Extremes | ESPmin* | Minimum Electrostatic Potential |
| 30 | | | ESPmax* | Maximum Electrostatic Potential |
| 31 | | Partitioned Molecular Surface Statistics | Overall_Surface_Area* | Overall Surface Area |
| 32 | | | Pos_Surface_Area* | Positive Surface Area |
| 33 | | | Neg_Surface_Area | Negative Surface Area |
| 34 | | | Overall_Average | Average Overall Electrostatic Potential |
| 35 | | | Pos_Average* | Average Positive Electrostatic Potential |
| 36 | | | Neg_Average | Average Negative Electrostatic Potential |
| 37 | | | Overall_Variance* | Variance of Overall Electrostatic Potential |
| | | | Nonpolar_Area | Nonpolar Area |

| | | | | |
|---|---|---|---|---|
| | | | Polar_Area* | Polar Area |
| 38 | | Integrated Polarity Indices | Nu* | Electrophilicity Index |
| 39 | | | Pi | Nucleophilicity Index |
| 40 | | | MPI* | Molecular Polarizability Index |
| 43 | Reactivity Indices | Averaged Local Ionization Energy | ALIEmin | Minimum Averaged Local Ionization Energy |
| 44 | | | ALIEmax | Maximum Averaged Local Ionization Energy |
| 45 | | | ALIE_Ave | Averaged Local Ionization Energy Average |
| 46 | | | ALIE_Var | Averaged Local Ionization Energy Variance |
| 47 | | | ALIEmin* | Minimum Local Electron Affinity |
| 48 | | Local Electron Affinity | LEAmax* | Maximum Local Electron Affinity |
| 49 | | | LEA_Ave* | Local Electron Affinity Average |
| 50 | | | LEA_Var* | Local Electron Affinity Variance |

**Table S3.** Names and corresponding literature DOIs for the 125 selected Y-series acceptor molecules.

| Number | Acceptor | DOI |
|---|---|---|
| 1 | BTP-BO-TBO | 10.1038/s41467-025-56799-6 |
| 2 | BTP-C11 | 10.1002/aenm.202100079 |
| 3 | BTP-ClBr1 | 10.1002/aenm.202002649 |
| 4 | BTP-DBO | 10.1038/s41467-025-56799-6 |
| 5 | BTP-DC11 | 10.1038/s41467-025-56799-6 |
| 6 | BTP-DTBO | 10.1038/s41467-025-56799-6 |
| 7 | BTP-H2 | 10.1039/d2ee00595f |
| 8 | BTP-S10 | 10.1002/aenm.202201076 |
| 9 | BTP-SA1 | 10.1002/advs.202405303 |
| 10 | BTP-T | 10.1002/adfm.202409723 |
| 11 | C9BO | 10.1002/aenm.202403121 |
| 12 | CB-2Se | 10.1002/adfm.202419176 |
| 13 | diDT-BO | 10.1002/adfm.202410786 |
| 14 | DM-F | 10.1002/ange.202407007 |
| 15 | DTC8 | 10.1038/s41467-025-56225-x |
| 16 | A-SSe-TCF | 10.1021/jacs.5c00004 |
| 17 | A-WSSe-Cl | 10.1002/anie.202104766 |
| 18 | AQx-0F | 10.1016/j.joule.2024.01.005 |
| 19 | AQx-1F | 10.1016/j.joule.2024.01.005 |
| 20 | AQx-1 | 10.1002/adma.201906324 |
| 21 | AQx-2F | 10.1016/j.joule.2024.01.005 |
| 22 | AQx-2 | 10.1002/adma.201906324 |
| 23 | AQx-3F | 10.1016/j.joule.2024.01.005 |
| 24 | BP4T-4F | 10.1002/aenm.202003177 |
| 25 | BPF-4F | 10.1021/acsenergylett.0c01688 |
| 26 | BPS-4F | 10.1021/acsenergylett.0c01688 |
| 27 | BPT-4F | 10.1021/acsenergylett.0c01688 |

| | | |
|---|---|---|
| 28 | BT-BO-L4F | 10.1021/jacs.1c00211.s001 |
| 29 | BT-LIC | 10.1021/jacs.1c00211.s001 |
| 30 | BTIC-CF3-γ | 10.1016/j.joule.2020.02.004 |
| 31 | BTP-2F-ThCl | 10.1016/j.joule.2020.03.023 |
| 32 | BTP-2ThCl | 10.1016/j.joule.2020.03.023 |
| 33 | BTP-4F-12 | 10.1002/adma.201903441 |
| 34 | BTP-4F-P2EH | 10.1002/aenm.202102596 |
| 35 | BTP-Bme | 10.1002/anie.202406153 |
| 36 | BTP-ClBr2 | 10.1002/aenm.202002649 |
| 37 | BTP-ClBr | 10.1002/aenm.202002649 |
| 38 | BTP-eC11 | 10.1002/adma.201908205 |
| 39 | BTP-eC7 | 10.1002/adma.201908205 |
| 40 | BTP-eC9 | 10.1002/adma.201908205 |
| 41 | BTP-PhC6 | 10.1039/d2ee01848a |
| 42 | BTP-S1 | 10.1002/adma.202001160 |
| 43 | BTP-S2 | 10.1002/adma.202001160 |
| 44 | BTP-S7 | 10.1038/s41467-021-24937-5 |
| 45 | BTP-S8 | 10.1038/s41467-021-24937-5 |
| 46 | BTP-S9 | 10.1038/s41467-021-24937-5 |
| 47 | BTP1O-4Cl-C10 | 10.1002/aenm.202003777 |
| 48 | BTP1O-4Cl-C12 | 10.1002/aenm.202003777 |
| 49 | BTP1O-4Cl-C8 | 10.1002/aenm.202003777 |
| 50 | DTY6 | 10.1016/j.joule.2020.07.028 |
| 51 | EHBzS-4F | 10.1021/acsenergylett.0c02230.s001 |
| 52 | L8-BO | 10.1038/s41560-021-00820-x |
| 53 | L8-HD | 10.1038/s41560-021-00820-x |
| 54 | L8-OD | 10.1038/s41560-021-00820-x |
| 55 | L8-ThCl | 10.1038/s41467-024-51359-w |
| 56 | m-BTP-PhC6 | 10.1021/acsenergylett.2c01364 |
| 57 | mBzS-4F | 10.1021/acsenergylett.0c02230.s001 |
| 58 | N-C11 | 10.1016/j.joule.2019.09.010 |
| 59 | N3-Cl-1 | 10.1039/d0ee02251a |
| 60 | N3-Cl-2 | 10.1039/d0ee02251a |
| 61 | N3 | 10.1039/d4ee01944j |
| 62 | o-BTP-PhC6 | 10.1039/d0ee03506h |
| 63 | SY2 | 10.1002/adfm.202000456 |
| 64 | Y18 | 10.1039/d0ee00862a |
| 65 | Y6-1O | 10.1002/aenm.202003141 |
| 66 | Y6 | 10.1016/j.joule.2019.01.004 |
| 67 | Z8 | 10.1038/s41560-024-01557-z |
| 68 | 2BO | 10.1002/aenm.202403121 |
| 69 | 6C-2F | 10.1039/d4ee02841d |
| 70 | 7C-2F | 10.1039/d4ee02841d |
| 71 | A-SSe-LSF | 10.1021/jacs.5c00004 |

| | | |
|---|---|---|
| 72 | A-SSe-4F | 10.1021/jacs.5c00004 |
| 73 | AA-1 | 10.1002/adma.202408858 |
| 74 | AA-2 | 10.1002/adma.202408858 |
| 75 | AQx-8 | 10.1002/aenm.202401561 |
| 76 | ATIC-C11 | 10.1002/adma.202413270 |
| 77 | B6Cl | 10.1021/jacs.4c01503.s001 |
| 78 | BO4Cl | 10.1021/acsenergylett.4c03168.s001 |
| 79 | BTA-C6 | 10.1038/s41467-024-55375-8 |
| 80 | BTA-E3 | 10.1038/s41467-024-55375-8 |
| 81 | BTA-E6 | 10.1038/s41467-024-55375-8 |
| 82 | BTA-E9 | 10.1038/s41467-024-55375-8 |
| 83 | BTA-HD-Rh | 10.1002/advs.202404997 |
| 84 | BTIC-CF3-m | 10.1016/j.joule.2020.02.004 |
| 85 | BTIC-CI-m | 10.1016/j.joule.2020.02.004 |
| 86 | BTIC-F-m | 10.1016/j.joule.2020.02.004 |
| 87 | BTP-2T | 10.1002/adfm.202409723 |
| 88 | BTP-4F-P3EH | 10.1002/aenm.202102596 |
| 89 | BTP-4F-PC6 | 10.1002/aenm.202102596 |
| 90 | BTP-4F | 10.1038/s41467-019-10351-5 |
| 91 | BTP-Biso | 10.1002/anie.202406153 |
| 92 | BTP-BO-TBO | 10.1038/s41467-025-56799-6 |
| 93 | BTP-Cy | 10.1002/adma.202418353 |
| 94 | BTP-eC9-4F | 10.1002/adfm.202417478 |
| 95 | BTP-OS | 10.1002/adfm.202415499 |
| 96 | BTP-PhC6-C11 | 10.1039/d2ee01848a |
| 97 | BTP-Ph | 10.1002/aenm.202100079 |
| 98 | BTP-T-BO | 10.1002/adfm.202409723 |
| 99 | BTP2O-4Cl-C12 | 10.1002/aenm.202003777 |
| 100 | C5-16 | 10.1002/adfm.202408340 |
| 101 | CB-Se | 10.1002/adfm.202419176 |
| 102 | DTC11-BO | 10.1002/adfm.202410786 |
| 103 | eC9-2Cl | 10.1002/adma.202102420 |
| 104 | eC9 | 10.1016/j.joule.2024.03.013 |
| 105 | K2 | 10.1002/adfm.202405168 |
| 106 | L8-BO-C4 | 10.1038/s41563-024-02087-5 |
| 107 | Qx-p-N4F | 10.1002/aenm.202403806 |
| 108 | SY1 | 10.1002/adfm.202000456 |
| 109 | TPT10 | 10.1021/jacs.9b09939.s001 |
| 110 | Y-BO-FCl | 10.1039/d1ee01832a |
| 111 | Y2CF3 | 10.1021/jacs.4c13471.s001 |
| 112 | Y7-BO | 10.1002/adfm.202200807 |
| 113 | AQx-22 | 10.1002/adma.202413376 |
| 114 | PzIC-SeSe-4F | 10.1002/adfm.202413259 |
| 115 | SMA | 10.1002/adma.202406690 |

| | | |
|---|---|---|
| 116 | BTP-B | 10.1002/adma.202418353 |
| 117 | AQx-21 | 10.1002/adma.202413376 |
| 118 | Y1 | 10.1002/adma.201904215 |
| 119 | Y2 | 10.1038/s41467-019-08386-9 |
| 120 | Y5 | 10.1002/adma.201807577 |
| 121 | 8C-2F | 10.1039/d4ee02841d |
| 122 | BT-L4F | 10.1021/jacs.1c00211.s001 |
| 123 | AQx-4F | 10.1016/j.joule.2024.01.005 |
| 124 | BTP-SA2 | 10.1002/advs.202405303 |
| 125 | BTP-SA3 | 10.1002/advs.202405303 |

**Table S4.** The contribution of different molecular substructures to the five descriptors and PCE obtained from SHAP analysis.

| Descriptors | Substructures | Substructures' contribution to descriptor | Substructures' weighted contribution to PCE |
|---|---|---|---|
| Pos_average | Terminal-group | 41.0% | 5.17% |
| | Central A′ unit | 33.2% | -4.2% |
| ESPmin | Terminal-group | 24.9% | 2.74% |
| | D unit | 16.4% | 1.81% |
| | Central A′ unit | 33.8% | -3.71% |
| Polar_area | Terminal-group | 24.2% | 2.13% |
| | D unit | 20.6% | -1.81% |
| MPI | Terminal-group | 32.0% | 2.24% |
| | D unit | 15.0% | -1.05% |
| ZZ | Side-chain | 39.7% | 2.65% |
| | Central A′ unit | 11.8% | -0.79% |
| | D unit | 17.3% | 1.16% |
| | Terminal-group | 15.1% | -1.01% |

## References


1. C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.* 2019, **15**, 1652-1671.
2. E. Caldeweyher, C. Bannwarth and S. Grimme, *J. Chem. Phys.* 2017, **147**, 034112.
3. F. Neese, F. Wennmohs, U. Becker and C. Riplinger, *J. Chem. Phys.* 2020, **152**, 224108.
4. F. Neese, *WIREs Computational Molecular Science* 2022, **12**.
5. A. D. Becke, *J. Chem. Phys.* 1993, **98**, 5648-5652.
6. C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B Condens. Matter.* 1988, **37**, 785-789.
7. F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.* 2005, **7**, 3297-3305.
8. G. Knizia, W. Li, S. Simon and H. J. Werner, *J. Chem. Theory Comput.* 2011, **7**, 2387-2398.
9. T. Lu and F. Chen, *J. Comput. Chem.* 2012, **33**, 580-592.