Wen *et al. J. Mater. Inf.* **2025**, *5*, 30
DOI: 10.20517/jmi.2024.102

Journal of Materials Informatics

Research Article

Open Access

Check for updates

Accelerated discovery of high-performance small-molecule hole transport materials via molecular splicing, high-throughput screening, and machine learning

Jiansen Wen[1,2,#], Shuwen Yang[1,3,4,#], Linqin Jiang[2,*], Yudong Shi[1], Zhihan Huang[4], Ping Li[2], Hao Xiong[2], Ze Yu[4], Xushan Zhao[4], Bo Xu[4], Bo Wu[1,3,*], Baisheng Sa[1,*], Yu Qiu[2]

[1]Multiscale Computational Materials Facility & Materials Genome Institute, School of Materials Science and Engineering, Fuzhou University, Fuzhou 350100, Fujian, China.
[2]Key Laboratory of Green Perovskites Application of Fujian Province Universities, Fujian Jiangxia University, Fuzhou 350100, Fujian, China.
[3]Materials Design and Manufacture Simulation Facility, School of Advanced Manufacturing, Fuzhou University, Jinjiang 362200, Fujian, China.
[4]Fujian Science and Technology Innovation Laboratory for Energy Devices (21C-Lab), Contemporary Amperex Technology Co., Limited (CATL), Ningde 352100, Fujian, China.
[#]Authors contributed equally.

[*]**Correspondence to:** Prof. Linqin Jiang, Key Laboratory of Green Perovskites Application of Fujian Province Universities, Fujian Jiangxia University, No. 2 Xiyuangong Road, Fuzhou 350100, Fujian, China, E-mail: linqinjiang@fjjxu.edu.cn; Prof. Bo Wu, Prof. Baisheng Sa, Multiscale Computational Materials Facility & Materials Genome Institute, School of Materials Science and Engineering, Fuzhou University, No. 2 Wulongjiang Avenue, Fuzhou 350100, Fujian, China, E-mail: wubo@fzu.edu.cn; bssa@fzu.edu.cn

How to cite this article: Wen, J.; Yang, S.; Jiang, L.; Shi, Y.; Huang, Z.; Li, P.; Xiong, H.; Yu, Z.; Zhao, X.; Xu, B.; Wu, B.; Sa, B.; Qiu, Y. Accelerated discovery of high-performance small-molecule hole transport materials via molecular splicing, high-throughput screening, and machine learning. *J. Mater. Inf.* **2025**, *5*, 30. https://dx.doi.org/10.20517/jmi.2024.102

Received: 30 Dec 2024 **First Decision:** 6 Feb 2025 **Revised:** 25 Mar 2025 **Accepted:** 2 Apr 2025 **Published:** 15 Apr 2025

Academic Editor: Sergei Manzhos **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

As the most representative and widely utilized hole transport material (HTM), spiro-OMeTAD encounters challenges including limited hole mobility, high production costs, and demanding synthesis conditions. These issues have a notable impact on the overall performance of perovskite solar cells (PSCs) based on spiro-OMeTAD and hinder its large-scale commercial application. Consequently, there exists a strong demand for high-throughput computational design of novel small-molecule HTMs (SM-HTMs) that are cost-effective, easy to synthesize, and offer excellent performance. In this study, a systematic and iterative design and development process for SM-

HTMs is proposed, aiming to accelerate the discovery and application of high-performance SM-HTMs. A custom-developed molecular splicing algorithm (MSA) generated a sample space of 200,000 intermediate molecules, culminating in the creation of a comprehensive database of over 7,000 potential SM-HTM candidates. In total, six promising HTM candidates were identified through MSA, density functional theory calculations and high-throughput screening. Furthermore, three machine learning algorithms, namely random forest, gradient boosting decision tree, and extreme gradient boosting (XGBoost), were employed to construct predictive models for key material properties, including hole reorganization energy, solvation free energy, maximum absorption wavelength, and hydrophobicity. Among these, the XGBoost-based model demonstrated the best overall performance. The MSA methodology combining comprehensive SM-HTM database and performance prediction models, as introduced in this study, offers a powerful and universal toolkit for the design and optimization of next-generation SM-HTMs, thereby paving the way for future advancements of PSCs.

Keywords: Perovskite solar cell, small-molecule hole transport materials, molecular splicing, density functional theory, high-throughput computational screening, machine learning

INTRODUCTION

With the growing urgency of addressing energy demands, climate change, and environmental challenges, the development of solar cells, which play a pivotal role in converting solar energy into electricity, has become increasingly significant[1-3]. Among various solar cell technologies, perovskite solar cells (PSCs) have emerged as a groundbreaking innovation. Due to their high power conversion efficiency (PCE), low cost, tunable optical bandgap, excellent carrier transport properties, broad spectral response, and straightforward fabrication processes, PSCs have rapidly become a focal point of research in the field of solar energy[4-7]. Notably, their PCE has reached an impressive value of 26.7%[8].

Hole transport materials (HTMs) are the essential component in PSCs for facilitating the transport of photo-generated holes to the electrodes, and the optimization of their performance can significantly improve the PCE of solar cells[9-11]. These materials can be broadly categorized into inorganic, polymeric and small-molecule HTMs (SM-HTMs)[9,12]. Inorganic HTMs, while promising, are limited by poor film-forming properties and a narrow range of material options, hindering their development. Polymer HTMs face limitations in their application scope due to the relatively intricate synthesis and purification processes, along with the challenges associated with accurately characterizing their molecular weight. On the other hand, SM-HTMs offer abundant availability and flexible structural tunability, enabling the design of target molecules with high hole mobility through various structural combinations[9,13]. Currently, spiro-OMeTAD is the most representative SM-HTM used in PSCs. It consists of two rigid π-conjugated systems connected by an orthogonal molecular conformation, offering excellent thermal stability and strong film-forming capabilities[14,15]. However, due to weak intermolecular interactions, spiro-OMeTAD films exhibit low hole mobility. To enhance device performance, dopants such as lithium salts and tBP are often introduced to improve conductivity and hole mobility. Unfortunately, the use of dopants can lead to device instability[16,17]. Moreover, the commercialization of spiro-OMeTAD is hindered by its complex synthesis, low yield, difficult purification, and high production cost[9]. Consequently, there is growing interest in the development of new SM-HTMs to overcome these limitations.

The function of the SM-HTM is to extract holes from the perovskite absorber layer and efficiently transport them to the electrode, thereby enhancing the performance of the solar cells. An ideal SM-HTM should meet the following criteria[9,18]: (i) its highest occupied molecular orbital (HOMO) energy level should match that of the perovskite layer, facilitating exciton separation at the interface, enabling easy hole injection into the hole transport layer (HTL), and preventing electron migration into the HTL; (ii) it should exhibit high hole

mobility, promoting rapid hole transport to the electrode; (iii) it should have good solubility and film-forming properties; (iv) it should possess good hydrophobicity, protecting the perovskite layer from moisture-induced degradation and improving device stability; (v) it should demonstrate high light transmittance, avoiding competition and absorption of light by the HTL that would otherwise be absorbed by the perovskite layer; (vi) it should be low-cost and easy to synthesize.

SM-HTMs with methoxyaniline as the terminal group have garnered significant attention from researchers due to their advantages of easy synthesis, straightforward purification, tunable energy levels, and strong structural functionality[12,19]. These materials typically consist of a central group and two terminal groups. Common central units include benzene rings, naphthalene, pyrrole, furan, thiophene, carbazole, and dibenzothiophene, while the terminal groups are often composed of dimethoxydiphenylamine and dimethoxytriphenylamine[20-23]. Studies have demonstrated that methoxyaniline groups play a critical role in influencing the electronic properties of HTMs. Additionally, the incorporation of methoxyaniline groups enhances the solubility of these materials, leading to improved film morphology[24-27].

Currently, research on new SM-HTMs primarily relies on traditional laboratory synthesis and trial-and-error methods, which are inefficient and costly. However, with the rapid advancement of artificial intelligence technology, high-throughput computational screening and machine learning (ML) methods have become increasingly valuable in materials science[28-30]. These approaches are now widely applied in the development of new materials for various fields, including photovoltaics, optoelectronics, and photocatalysis[31-34]. Therefore, combining high-throughput computational screening with ML to design new SM-HTMs and predict their properties represents a highly promising direction for future research and development of SM-HTMs. In recent years, significant progress has been made in the structural design and performance investigation of HTMs for solar cells from the perspective of theoretical chemistry[35-39]. Building on this foundation, the integration of high-throughput screening and ML has emerged as a pivotal driving force in accelerating the discovery of novel SM-HTMs. Wu *et al.* pioneered a closed-loop workflow combining high-throughput organic synthesis with Bayesian optimization (BO) to discover SM-HTMs tailored for perovskite devices[40]. By training predictive models on 149 synthesized molecules and screening a virtual library of 1 million candidates, they achieved a certified PCE of 25.9%, demonstrating the power of data-driven approaches in navigating complex material landscapes with limited datasets. Complementing this, Faruque *et al.* employed a translational dimer model for high-throughput screening of 74 diacenaphtho-extended heterocycles, coupled with ML-guided crystal structure prediction (CSP) and carrier mobility calculations[41]. Their workflow identified candidates with hole mobilities exceeding 10 $cm^2/V·s$, validated by semiclassical Marcus theory, highlighting the role of computational screening in optimizing molecular packing and charge transport. These studies exemplify a paradigm shift toward autonomous, ML-enhanced material discovery, offering scalable strategies for designing next-generation HTMs with superior optoelectronic properties. However, the success of high-throughput screening and ML hinges on the availability of a robust library of candidate structures for HTMs. Consequently, the development of innovative design methodologies and algorithms for the structure of SM-HTMs is essential.

In this study, an efficient design strategy for SM-HTMs is proposed, integrating molecular assembly algorithms, high-throughput screening, and ML model predictions. A self-developed molecular splicing algorithm (MSA) was employed to construct a database of potential SM-HTMs for PSCs. By integrating density functional theory (DFT), high-throughput computations were conducted to identify six high-performance candidate SM-HTMs for subsequent synthesis and performance evaluation. Furthermore, the molecular structure and property datasets obtained through high-throughput calculations were utilized to develop property prediction models for SM-HTMs using three ML approaches: random forest (RF),

gradient boosted decision trees (GBDTs), and extreme gradient boosting (XGBoost). The results indicate that the model built with the XGBoost algorithm not only requires the least training time but also delivers the best prediction accuracy, demonstrating the most comprehensive performance. This work will offer a rapid, efficient and cost-effective approach for the development of new SM-HTMs by combining MSA with high-throughput computational screening and ML.

MATERIALS AND METHODS

First principle calculation

All DFT calculations were conducted using Gaussian 16 software[42]. Initial optimization was carried out using the semi-empirical AM1 method during the molecular splicing stage[43]. Under gas-phase conditions, geometry optimization, HOMO energy levels, hole reorganization energy, and absorption spectra were calculated at the B3LYP level with the 6-31++G(d,p) basis set[44]. Furthermore, the solvation free energy of the solvent molecules in n-octanol and water was calculated using the M062X functional and the 6-31++G(d,p) basis set to estimate the hydrophobicity (LogP)[45,46].

Based on the optimized structure, the HOMO level, hole reorganization energy, solvation free energy, maximum light absorption wavelength, hydrophobicity, and synthetic feasibility score (SAScore) of the molecule were calculated. Hole reorganization energy is a key parameter for calculating hole mobility based on Marcus theory, which is given as follows[47]:

$$k_{\mathrm{h}}=\frac{4\pi}{\hbar}v^{2}\frac{1}{\sqrt{4\pi\lambda k_{\mathrm{B}}T}}\exp\left(\frac{-\lambda}{4k_{\mathrm{B}}T}\right) \tag{1}$$

where $\hbar$ is Planck's constant, v is the transfer integral, λ is the hole reorganization energy, k_{B} is the Boltzmann constant, and T is the Kelvin temperature. The smaller the hole reorganization energy, the higher the hole mobility. λ is calculated by[48,49]

$$\lambda=\lambda_{0}+\lambda_{+}=\left(E_{0}^{*}-E_{0}\right)+\left(E_{+}^{*}-E_{+}\right) \tag{2}$$

where λ_{0} represents the energy difference between different neutral state structures, λ_{+} is the energy difference between different cationic state structures, E_{0} is the energy obtained after optimizing the neutral molecular structure, and E_{+}^{*} represents the cationic energy under the geometric configuration of the neutral molecule. E_{0}^{*} is the energy of the neutral molecule under the cationic geometry, and E_{+} represents the energy obtained after optimizing the cationic structure. The solvation free energy ΔG_{solv} refers to the change in the free energy of the solute as it transitions from the gaseous state to the solution, which is given by[50]

$$\Delta G_{\mathrm{solv}}=E_{\mathrm{SMD}}-E_{\mathrm{gas}} \tag{3}$$

where E_{SMD} is the single-point energy under the solvation model for density (SMD) model, and the E_{gas} represents the single-point energy under the gas phase. The smaller the solvation free energy, the stronger the solubility of the solute in the chlorobenzene solvent. HTMs must exhibit good hydrophobicity to protect the perovskite layer from water vapor degradation and enhance the stability and lifespan of the device. Hydrophobicity can be quantified using the n-octanol-water partition coefficient and LogP, which is calculated as follows[51-53]:

$$\mathrm{Log}P = \frac{\left(\Delta G_{\mathrm{oct}} - \Delta G_{\mathrm{w}}\right)}{2.303RT} \tag{4}$$

where ΔG_{oct} is the free energy of the molecule in n-octanol, ΔG_w is the free energy of the molecule in water, and R is the standard molar gas constant. The SAScore is a rapid metric used to assess the synthesis difficulty of a molecule. The score ranges from 1 to 10, with values closer to 1 indicating easier synthesis and values closer to 10 indicating greater difficulty, which is given as follows[54]:

$$\mathrm{SAScore} = \mathrm{FragmentScore} - \mathrm{ComplexityPenalty} \tag{5}$$

The FragmentScore is calculated based on 1 million representative molecules selected from the PubChem database. The ComplexityPenalty is a composite score that accounts for the presence of non-standard structures in the molecule, such as large rings, non-standard ring structures, and three-dimensional complex architectures. The SAScore is used as an evaluation metric for HTMs, providing a preliminary assessment of the synthesis difficulty of target molecules.

ML

All ML models were implemented using the scikit-learn[55] and xgboost[56] packages. In this study, the RDkit toolkit was used to extract 208 molecular descriptors from the structural data, including 12 basic descriptors (e.g., molecular weight, valence electron count), 38 descriptors related to molecular surface area (MolSurf), 19 topological chemical descriptors (GraphDescriptors), and 85 molecular fragment descriptors. These included eight two-dimensional descriptors (BCUT2D), one drug-like descriptor (QED), two Crippen descriptors, 18 Lipinski descriptors, and 25 electrotopological state index descriptors (Estate). The training-to-test set ratio was 9:1, and the model hyperparameters were optimized using grid search with 10-fold cross-validation. Data normalization was applied to prevent gradient instability and overfitting, thereby improving the model's accuracy and convergence speed. The performance of the regression models was evaluated using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2), as given below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{7}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{8}$$

where N is the total number of samples, y_i is the true value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the average of the true value y_i. The larger the RMSE and MAE values, the poorer the model performance, and conversely, the smaller these values, the better the model performance. The closer the R^2 value is to 1, the better the model fit.

RESULTS AND DISCUSSION

D-π-D molecular splicing design

The target structure for the design of the HTM molecule is the D-π-D type, which features a stable planar structure, where D represents the dimethoxydiphenylamine group and π is the intermediate. Figure 1 illustrates the complete process of D-π-D molecular splicing design, which is primarily divided into two stages: the splicing of the intermediate π molecule and the splicing of the end group D with the intermediate π molecule.

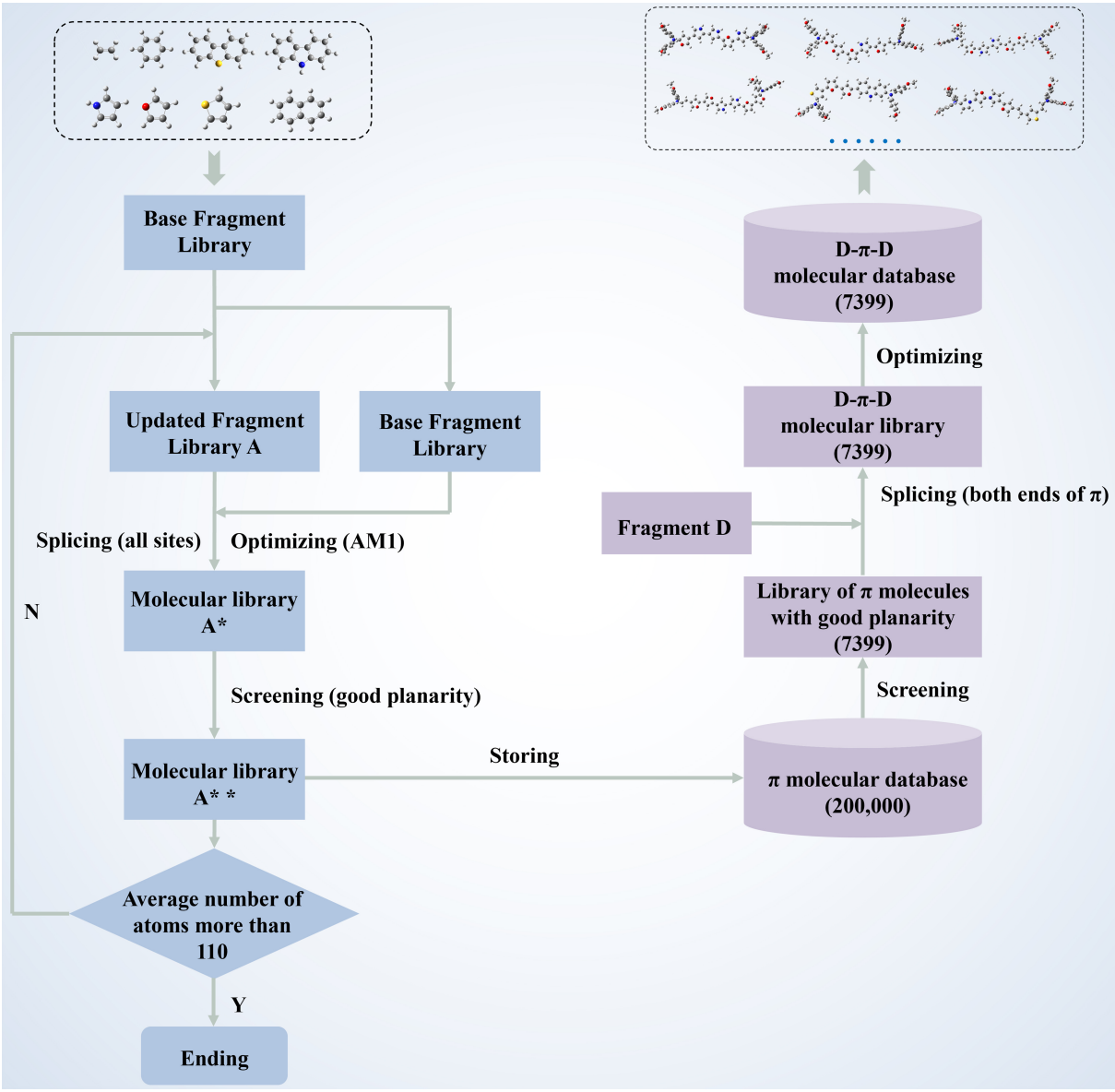

Figure 1. Schematic diagram of the molecular splicing workflow for designing D-π-D molecules.

Fragment-based molecular splicing design of the π structures

The rules for designing the intermediate π splicing are as follows: (1) the structure should be as conjugated as possible, with high molecular planarity; (2) the number of atoms in the molecule should range between 100 and 200. The library of basic molecular fragments includes ethylene, benzene rings, naphthalene, pyrrole, furan, thiophene, carbazole, and dibenzothiophene, as these molecules exhibit excellent charge transport properties due to their intermolecular interactions. Using these eight molecular fragments as the basic backbone for molecular splicing, the intermediate π molecular splicing process is shown in Figure 1.

The fragment splicing algorithm developed for π molecules follows a strategy where the molecular library A is updated in each iteration, while the basic molecular library B remains fixed. This approach is analogous to performing molecular fragment growth on the molecules in library A during each iteration. Generally, the process consists of the following four steps. (i) Identification of Splicing Sites: Identify all potential splicing

positions on the molecule. The potential splicing sites correspond to the positions of hydrogen atoms on the molecular fragments. Splicing between molecular fragments involves substituting the hydrogen atoms at these specific positions. To identify all possible splicing sites on a molecule, it is necessary to scan through all atoms in the molecular fragments, locate each hydrogen atom, and assign a unique identifier to each, which will then be used for subsequent fragment splicing. Additionally, the symmetry of the molecular structure should be taken into account, as symmetric sites need only be considered once, thus reducing the computational cost in the subsequent splicing process; (ii) Combination and Structural Optimization: Combine the molecules in library A with the basic fragment library B, and perform batch structural optimization using the semi-empirical quantum calculation method AM1 in Gaussian16 software. The optimized molecular structures are then stored in the molecular library A^*; (iii) Structural Screening: Conduct structural screening on the molecules in library A^* after each splicing round, retaining those with good planarity and storing them in the molecular library A^{**}. This step significantly reduces computational load and storage space while ensuring that the molecules in the initial library A of each round exhibit good planarity, thereby increasing the likelihood of obtaining molecules with desirable planarity in subsequent rounds. The molecules from the molecular library A^{**} generated in each iteration are systematically archived in the final π molecular database; (iv) Atom Count Evaluation: Evaluate the atom count of the molecules in the A^{**} library. If the atom count is less than 110, the splicing process is repeated. If the count exceeds 110, the π molecular splicing process concludes. After 19 rounds of splicing, 200,000 π molecular structures were generated in the π molecular database.

Molecular splicing design of the D-π-D structures

Upon obtaining the π molecular structure database, the subsequent step is to attach D fragments to both ends of the intermediate π molecules, thereby forming the target D-π-D structures. The D fragments are spliced at the most distant positions on the π molecules. The process is carried out in three main steps. (1) Planarity Screening: A planarity screening is conducted on the 200,000 structures in the π molecular database, resulting in the selection of 7,399 molecules with excellent planarity; (2) Molecular Splicing: D fragments are attached to both ends of the selected π molecules to form the complete D-π-D target molecular structures. Given that multiple potential splicing sites exist on the π molecule, the two most distant sites are selected as the connection points for the D fragment. The RDkit toolkit is used to convert the molecular structure into a graph-based representation, from which a distance matrix is computed. By examining the elements of this matrix, the two nodes with the largest separation are identified as the splicing sites for the D fragment; (3) Structural Optimization: The resulting D-π-D molecules undergo structural optimization, and the optimized structures of the 7,399 molecules are stored in the D-π-D molecular structure database for subsequent high-throughput property calculations. Four representative D-π-D spliced molecules with varying atom counts are presented in Supplementary Figure 1.

High-throughput computational screening

High-throughput calculations were performed on 7,399 selected molecules in the D-π-D molecular database to determine their HOMO energy levels, hole reorganization energies, solvation free energies, maximum absorption peaks, and hydrophobicity (LogP) values. Among these, 7,222 molecules yielded normal results from DFT calculations, while 177 computational tasks failed to converge. Finally, the SAScore was computed for the 7,222 successfully converged molecules using the RDKit toolkit. Specific information on these molecules, including molecular structures and calculated properties, is summarized in Supplementary Files 1 and 2, respectively.

Figure 2 illustrates the high-throughput computational results for various performance parameters of 7,222 D-π-D HTM molecules. The x-axis represents the ID number of the molecule, with larger numbers corresponding to molecules with a greater number of atoms. As shown in Figure 2A and B, no significant correlation is observed between the HOMO energy levels or the hole reorganization energies and the

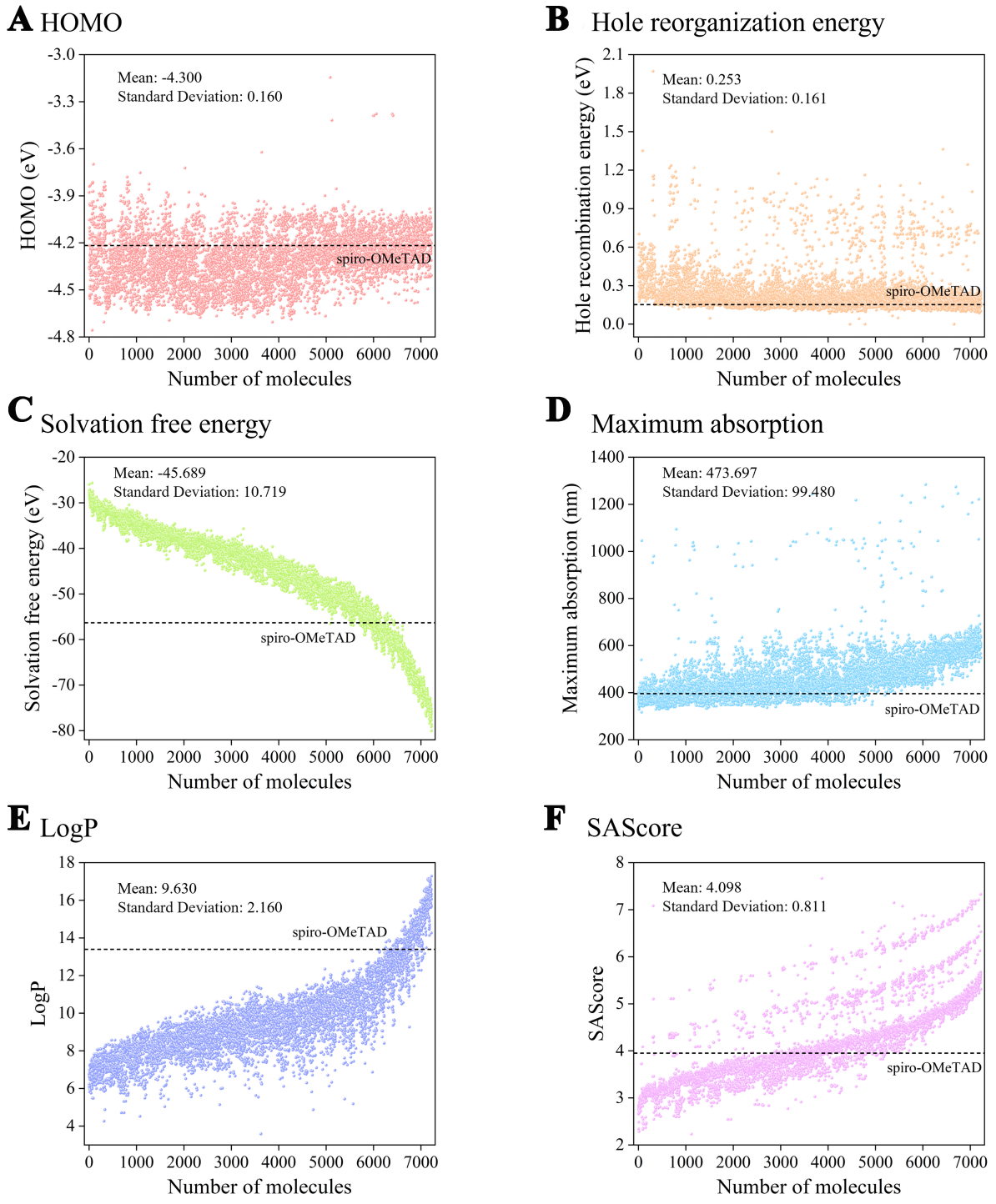

Figure 2. High-throughput computational results for various performance parameters of 7,222 D-π-D HTM molecules: (A) HOMO, (B) Hole reorganization energy, (C) Solvation free energy, (D) Maximum absorption, (E) LogP, and (F) SAScore, respectively. The x-axis of the images (A) to (F) represents the ID number of the 7,222 D-π-D HTM molecules, with larger numbers corresponding to molecules with a greater number of atoms. The dashed lines in the figure represent the calculated performance parameters of spiro-OMeTAD, serving as reference values. HTM: Hole transport material; HOMO: highest occupied molecular orbital.

number of atoms. However, the solvation free energy exhibits an inverse relationship with the number of atoms, meaning that larger molecules tend to have lower solvation free energies, which enhances their solubility in chlorobenzene solvent, as depicted in Figure 2C. Chlorobenzene was selected as the solvent for solvation free energy calculations due to its widespread use in PSC fabrication, particularly for dissolving SM-HTMs such as spiro-OMeTAD[57-60]. This choice ensures that the solvation free energy calculations align with experimental conditions, providing a reliable basis for predicting HTM performance and facilitating comparisons with spiro-OMeTAD. As a weakly polar solvent, chlorobenzene effectively dissolves non-polar or weakly polar HTMs, preventing excessively strong solvent-molecule interactions. This ensures uniform HTL film formation and minimizes over-solvation or interference with the crystallization of the perovskite layer, ultimately improving the device's optoelectronic efficiency and stability. The choice of solvent is crucial in determining the solvation free energy and solubility of HTM molecules. Strong polar solvents [such as N,N-dimethylformamide (DMF) or ethanol] may reduce solvation free energy and increase solubility, but they can also leave solvent residues or damage the perovskite layer. Non-polar solvents (such as toluene or n-hexane) may increase solvation free energy and decrease solubility. Chlorobenzene, as a weakly polar solvent, strikes an optimal balance between solubility and film formation, making it suitable for processing most HTMs. Figure 2D reveals that the maximum absorption peak increases with the number of atoms, indicating a redshift in the absorption spectrum as molecular size grows. Additionally, as illustrated in Figure 2E, the hydrophobicity (LogP value) is directly proportional to the number of atoms, implying that larger molecules exhibit stronger hydrophobic characteristics. Finally, the SAScore of a molecule also increases with the number of atoms, as shown in Figure 2F, indicating that larger molecules are more challenging to synthesize.

In addition, performance data for seven previously reported HTMs, including spiro-OMeTAD[61], DTPC8-ThDTPA[61], DTPC13-ThTPA[62], DTP-C6Th[63], YZ18[64], YZ22[64], and TPA-TVT-TPA[65], were also calculated, with detailed values provided in Supplementary Table 1. It is important to note that the B3LYP functional tends to underestimate the HOMO-LUMO gap and overestimate the HOMO level. As a result, the calculated HOMO values are generally higher than experimental values, consistent with prior literature[61-63,66,67]. This alignment validates the accuracy of our structural model and computational methodology. It is noteworthy that this study is designed to address the critical limitations of spiro-OMeTAD, including its low hole mobility, demanding synthesis requirements, and high production costs, through the development of novel SM-HTMs. The performance data of spiro-OMeTAD were used as a reference for screening high-throughput calculation results.

The screening process was primarily based on six criteria: HOMO level, hole reorganization energy, solvation free energy, maximum absorption peak, LogP, and SAScore. First, 7,222 molecules were pre-screened based on the HOMO energy level and maximum absorption peak. These two parameters are critical for identifying candidate materials capable of effectively replacing spiro-OMeTAD in photovoltaic devices. Specifically, the HOMO energy level ensures proper energy alignment between the HTL and the perovskite absorber, which is essential for efficient charge extraction. Simultaneously, the absorption maximum mitigates spectral overlap with the perovskite layer, thereby minimizing parasitic light absorption by the HTL and maximizing light utilization by the perovskite active layer. These combined properties guarantee that the newly designed SM-HTMs can be seamlessly integrated into device architectures analogous to those employing spiro-OMeTAD. The calculated HOMO value of spiro-OMeTAD ± 0.1 eV (i.e., -4.117 to -4.317 eV) was used as the criterion for HOMO screening. Furthermore, referencing spiro-OMeTAD, we also set a screening criterion for the maximum absorption peak. To prevent excessive light absorption by the HTL, which could compromise device efficiency, the maximum absorption peak of the 7,222 molecules was required to be less than 400 nm, ensuring optimal photoelectric conversion efficiency.

Next, the remaining molecules were ranked according to their comprehensive properties based on four metrics: hole reorganization energy, solvation free energy, LogP, and SAScore. Given the significant variations in value ranges for these metrics, Min-Max normalization was applied to scale all indicators to a range of [0, 1]. For metrics where lower values indicate better properties (hole reorganization energy, solvation free energy, and SAScore), reverse normalization was applied. Furthermore, to address the limitations of spiro-OMeTAD, such as low hole mobility and high synthesis difficulty, higher weights were assigned to these properties. Specifically, weights of 0.3 were assigned to hole reorganization energy and SAScore, while solvation free energy and LogP were assigned weights of 0.2. The preliminarily screened molecules were scored and ranked based on these weighted criteria. A flowchart of the process is shown in Figure 3A, and the scores of the top 20 molecules are shown in Supplementary Table 2. The 6 highest-scoring candidate molecules are screened out, denoted as M1-M6, which are presented in Figure 3B. To assess the robustness of the weighting scheme, we systematically varied the weights of the four properties (hole reorganization energy, solvation free energy, LogP, and SAScore) within a range of 0.1 to 0.4 using a step size of 0.01 (maintaining the total sum of weights equal to 1), generating a total of 124 unique weighting combinations. For each combination, the weighted scores of all candidate molecules were calculated, and the distribution of the top 10 molecules with the highest scores is shown in Supplementary Figure 2A. The molecular IDs and their occurrence frequencies are presented in Supplementary Table 3 and Supplementary Figure 2B. Remarkably, 50% of the most frequently occurring top ten molecules across all weighting scenarios (highlighted as red bars in Supplementary Figure 2B) overlap with our final selection, demonstrating the stability of our weighted screening strategy.

The molecular properties of the six screened SM-HTM molecules M1-M6 are listed in Supplementary Table 4. Figure 4 compares the performance index data of the filtered M1-M6 molecules with that of several common SM-HTMs, such as spiro-OMeTAD, DTPC8-ThDTPA, DTPC13-ThTPA, DTP-C6Th, YZ18, YZ22, and TPA-TVT-TPA. The performance of the M1-M6 molecules remains promising when compared to spiro-OMeTAD and several other HTMs. The HOMO levels of the six screened molecules are very close to that of spiro-OMeTAD and are all higher than the common valence band energy level of perovskite materials. This theoretically ensures that the perovskite material absorbs incident photons to form electron-hole pairs, with the holes being able to enter the HTL more readily. The maximum light absorption peaks of these molecules are all below 400 nm, avoiding overlap with the visible light absorption range of perovskite materials. This ensures that M1-M6 molecules will not compete for light with the perovskite layer, allowing perovskite materials to absorb and utilize sunlight more efficiently. In terms of hole reorganization energy, as depicted in Figure 4A, the 6 molecules are similar to spiro-OMeTAD overall and show significant improvement over YZ18 and YZ22, indicating that their hole mobility remains substantially excellent. Figure 4B reveals that the solvation free energies of the 6 molecules are comparable to those of commonly used HTMs, indicating their favorable solubility in chlorobenzene solution. As illustrated in Figure 4C, the LogP values of the screened molecules are maintained around 10, significantly higher than that of YZ22, indicating their superior hydrophobic properties. Furthermore, the SAScores of these molecules demonstrate a significant improvement over YZ18, YZ22, and TPA-TVT-TPA, suggesting their promising synthetic prospects, as shown in Figure 4D.

Although the screened SM-HTMs have not yet been experimentally synthesized and validated, their synthetic accessibility is strongly supported by the design principles of previously reported and successfully synthesized SM-HTMs[20-22,38], combined with the use of the SAScore parameter in our screening process, a metric specifically designed to evaluate synthetic feasibility. Furthermore, the stringent screening based on critical performance parameters, including HOMO energy levels, hole reorganization energy, solvation free energy, maximum absorption wavelength, and hydrophobicity, ensures their practical applicability. This

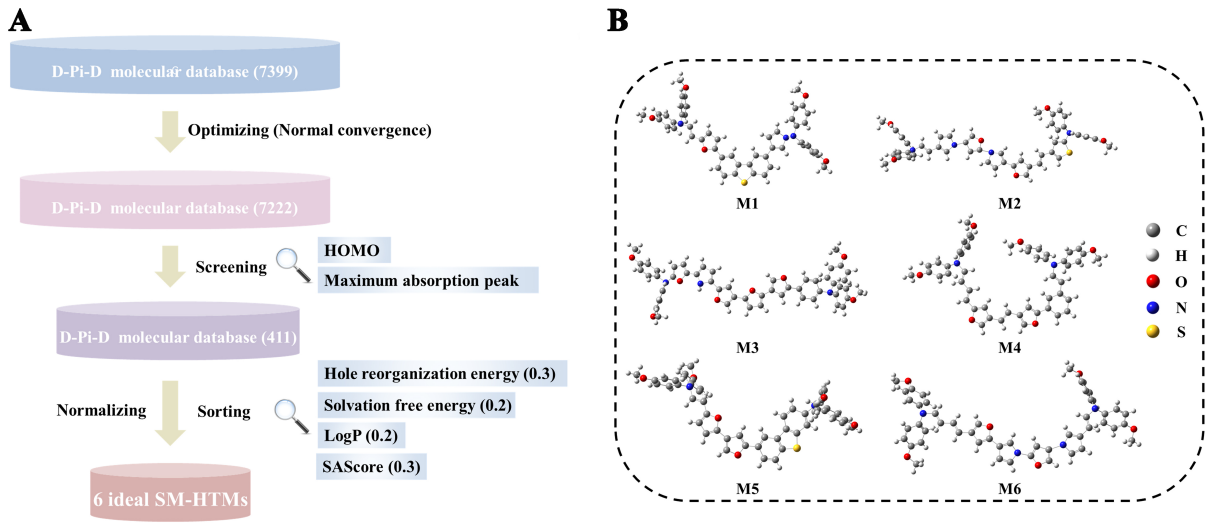

Figure 3. (A) Schematic diagram of the screening process for SM-HTMs; (B) The structures of 6 highest-scoring candidate SM-HTM molecules screened by high-throughput. SM-HTMs: Small-molecule hole transport materials.

comprehensive evaluation framework enhances the likelihood that the selected molecules are not only synthesizable but also functionally viable for real-world applications. The computational predictions presented here provide a robust foundation to streamline subsequent synthetic efforts, thereby guiding and accelerating experimental validation.

The synthesis complexity, economic feasibility, and environmental impact of HTMs are significantly influenced by their molecular structures. Linear SM-HTMs designed and screened in this study generally exhibit lower synthetic complexity due to well-established coupling or condensation reactions, higher yields, and simpler purification processes, making them more suitable for high-throughput screening and large-scale production. Their relatively straightforward synthesis also translates to lower production costs and reduced solvent and catalyst consumption, thereby minimizing their environmental footprint. Therefore, the linear SM-HTMs identified in this study hold significant potential for future applications in the photovoltaic field and are expected to serve as viable alternatives to traditional HTMs, enhancing device efficiency and processability while maintaining cost advantages and promoting sustainability.

Establishment and validation of ML predictive models

A database of 7,222 molecular structure-performance data was constructed through the aforementioned high-throughput calculations. Subsequently, RF, GBDT, and XGBoost methods were used to build molecular structure-property models for four properties in the database: hole reorganization energy, maximum light absorption peak, hydrophobicity, and solvation free energy. The performance and predictive effectiveness of the models were then evaluated and analyzed.

Establishment of ML predictive models

Figures 5-7 present the true values and ML predictions for hole reorganization energy, solvation free energy, maximum absorption, and LogP, based on the RF, GBDT, and XGBoost models, respectively. The specific performance indicators R^2, MAE, and RMSE are listed in Supplementary Table 5. Additionally, the 15 most important descriptors for each property (hole reorganization energy, solvation free energy, maximum absorption, and LogP) identified through the RF, GBDT, and XGBoost models are shown in Supplementary Figures 3-5, with detailed explanations of the relevant features provided in Supplementary Tables 6-17. As

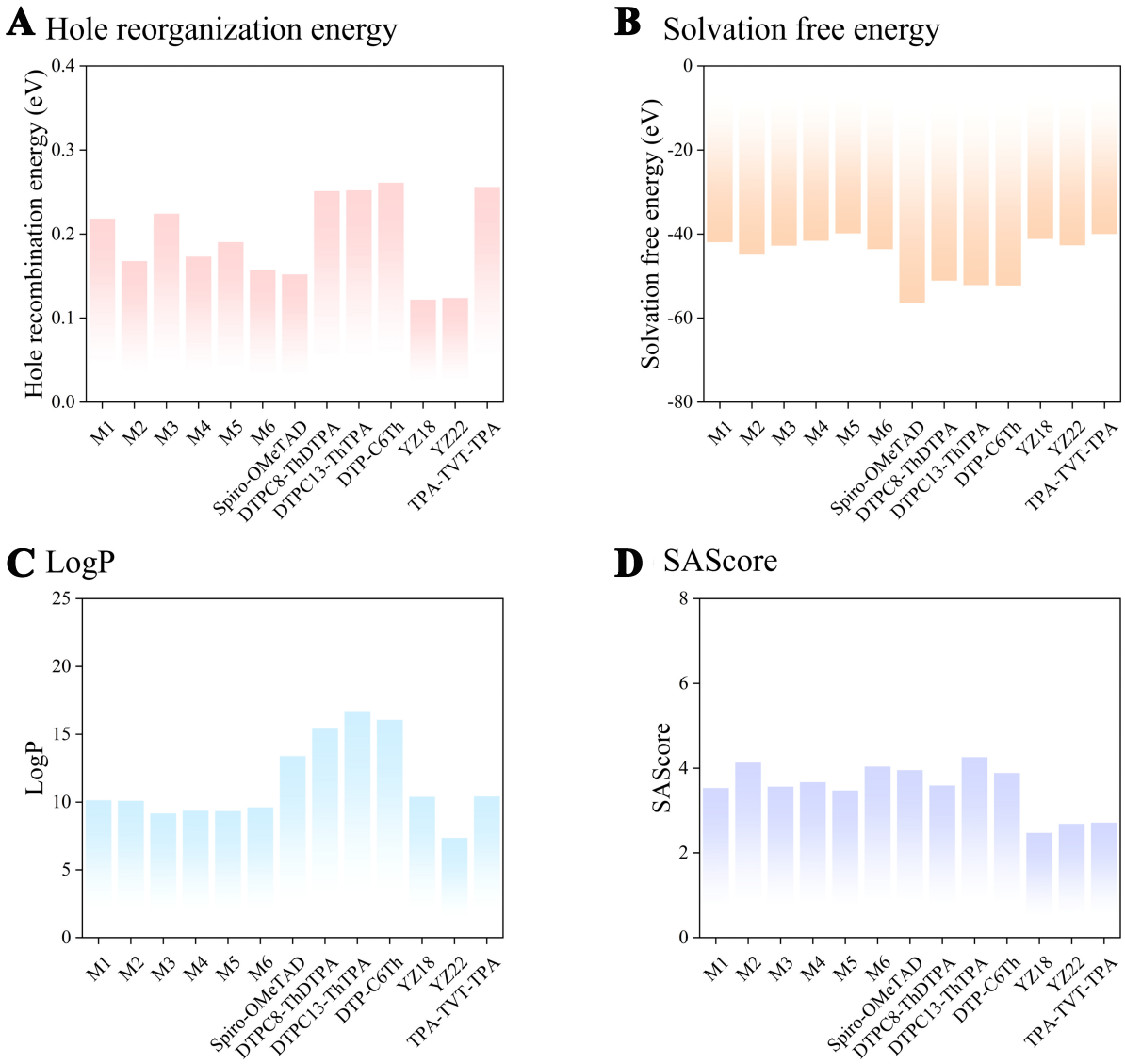

Figure 4. Comparison of (A) Hole reorganization energy, (B) Solvation free energy, (C) LogP, and (D) SAScore of the six selected molecules and several common SM-HTMs, respectively. SM-HTMs: Small-molecule hole transport materials.

shown in Figure 5, the RF model demonstrates excellent prediction capabilities for solvation free energy, maximum light absorption peak, and hydrophobicity, but performs moderately for hole reorganization energy, with the R^2 value for the test set reaching only 0.739. In contrast, the GBDT model excels in predicting maximum light absorption and hydrophobicity, as well as solvation free energy, and shows a notable improvement in predicting hole reorganization energy, with an R^2 value of 0.865, compared to the RF model's performance. Notably, the XGBoost model delivers superior performance across all four properties. Specifically, for hole reorganization energy, the R^2 value reaches 0.901, a significant improvement over the RF (0.739) and GBDT (0.865) models. Furthermore, the prediction performance (R^2 values) of the RF, GBDT, and XGBoost models across the four datasets (hole reorganization energy, solvation free energy, maximum light absorption, and hydrophobicity) are ranked as follows: XGBoost (0.901) > GBDT (0.865) > RF(0.739); solvation free energy: XGBoost (0.998) > GBDT (0.997) > RF (0.996); maximum absorption peak: XGBoost (0.969) > GBDT (0.946) > RF (0.923); hydrophobicity: XGBoost

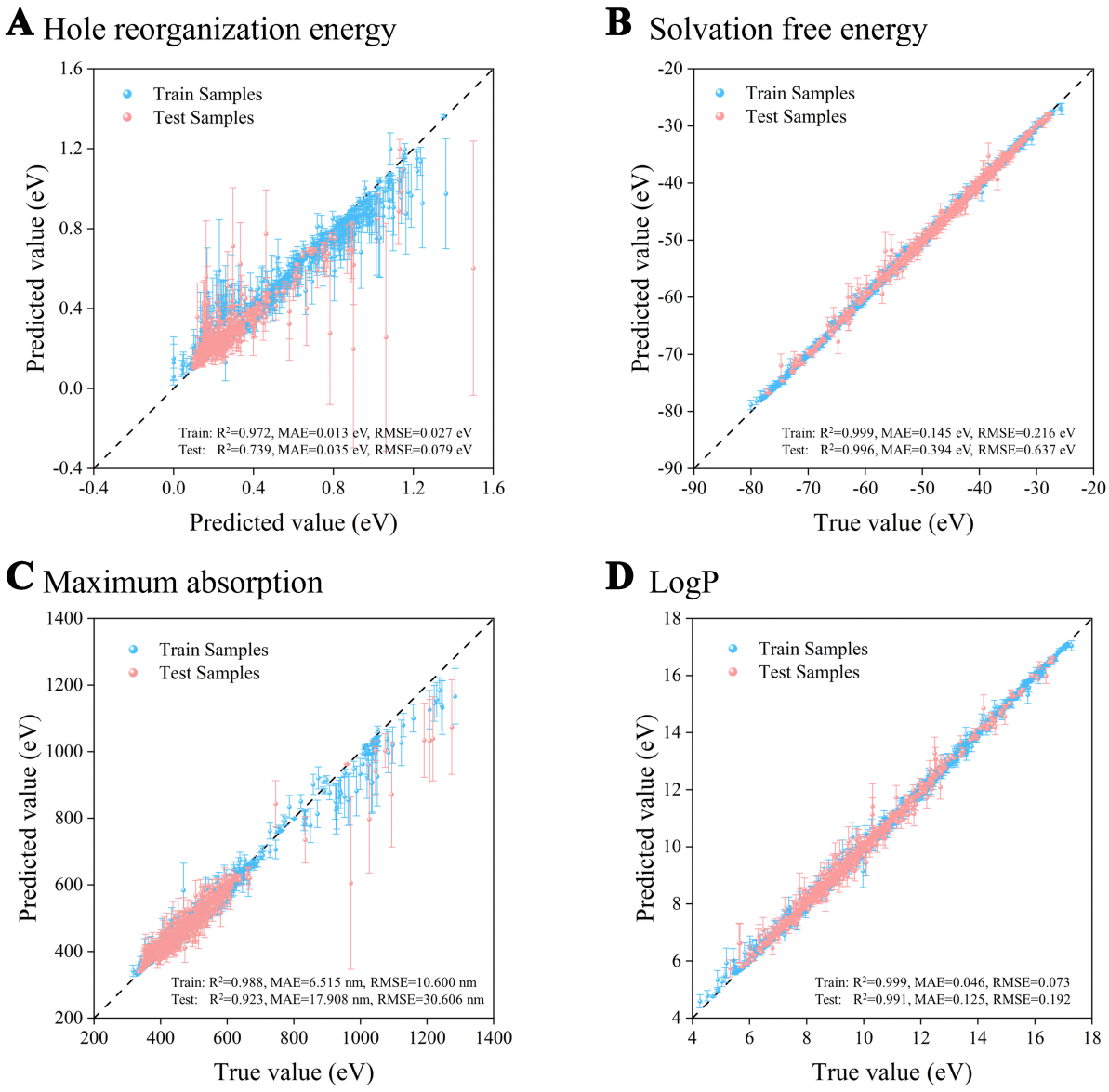

Figure 5. The true values and ML predicted values for (A) Hole reorganization energy, (B) Solvation free energy, (C) Maximum absorption, and (D) LogP based on the RF model, respectively. ML: Machine learning; RF: random forest.

(0.996) > GBDT (0.992) > RF (0.991). It is evident that the XGBoost model outperforms all others in predicting all four datasets. In terms of computational efficiency, the average training time for the three models is as follows: XGBoost (8.2 s) < GBDT (206.3 s) < RF (611.3 s), highlighting the significantly higher computational efficiency of the XGBoost model compared to GBDT and RF. Overall, the XGBoost model demonstrates the best performance across the current dataset, with excellent generalization and computational efficiency. The superior predictive accuracy and computational efficiency of XGBoost can be attributed to its advanced algorithmic design. First, it employs an optimized gradient-boosted framework integrated with L1/L2 regularization techniques to mitigate overfitting and enhance generalization capabilities. Unlike RF, which relies on averaging multiple decision trees and may inadequately capture complex feature interactions, XGBoost effectively models intricate relationships through its sequential tree-building process. Second, it leverages a weighted quantile sketch for sparse data optimization and

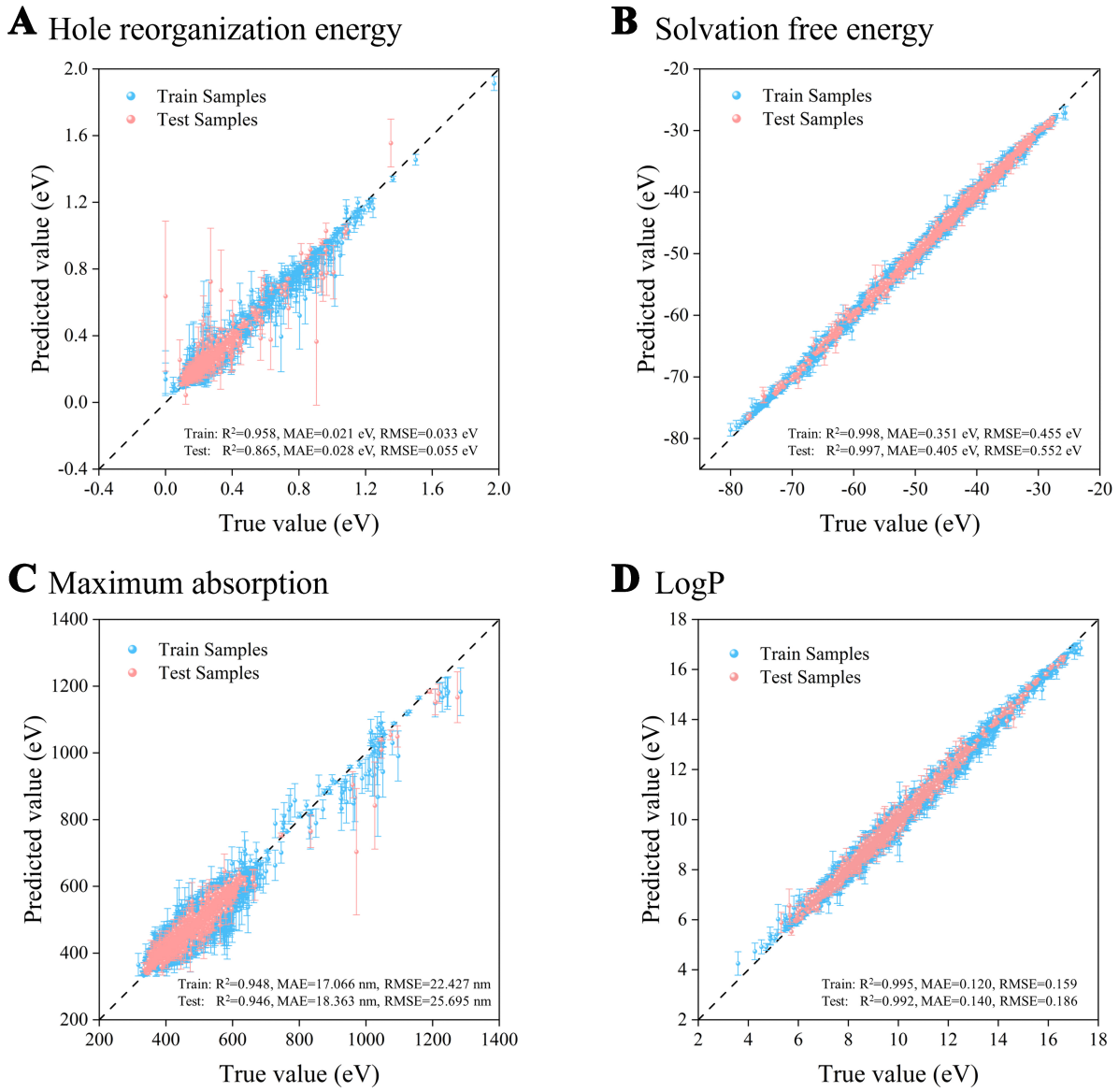

Figure 6. The true values and ML predicted values for (A) Hole reorganization energy, (B) Solvation free energy, (C) Maximum absorption, and (D) LogP based on the GBDT model, respectively. ML: Machine learning; GBDT: gradient boosted decision tree.

parallelized tree learning, significantly accelerating computational speed while maintaining precision. In contrast, conventional GBDT lacks such systematic optimizations, resulting in suboptimal efficiency-accuracy trade-offs. These innovations collectively enable XGBoost to achieve a balanced and robust performance in both accuracy and scalability.

Validation of ML predictive models

To further assess the performance of the model on a new dataset, seven reported HTMs, including spiro-OMeTAD, DTPC8-ThDTPA, DTPC13-ThTPA, DTP-C6Th, TPA-TVT-TPA, YZ18, and YZ22, were selected as test samples. The RF, GBDT, and XGBoost ML models were then employed to predict the hole reorganization energy, solvation free energy, maximum light absorption peak, and hydrophobicity of these molecules. The generalization ability of the models was evaluated by comparing the predicted values from

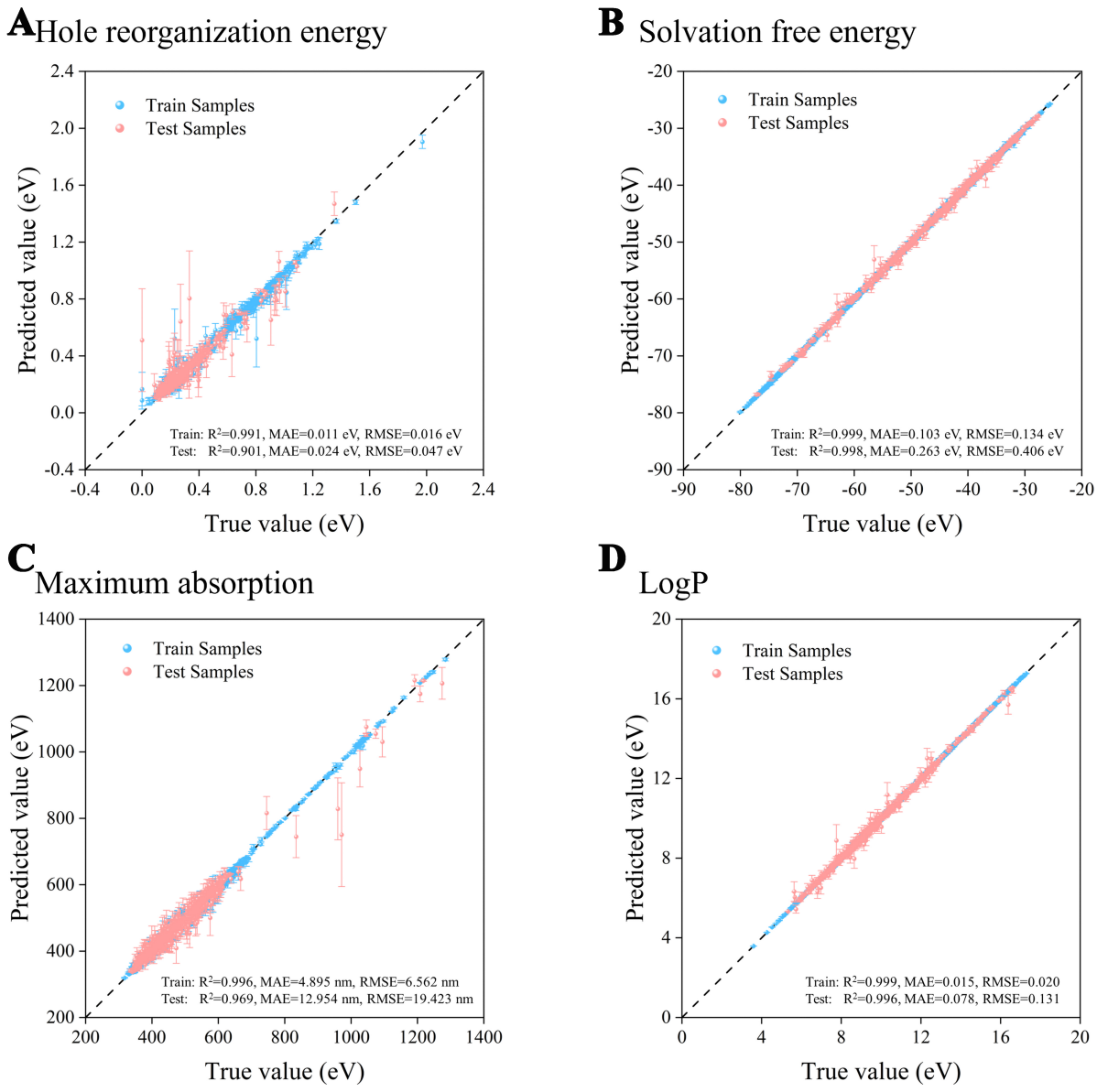

Figure 7. The true values and ML predicted values for (A) Hole reorganization energy, (B) Solvation free energy, (C) Maximum absorption, and (D) LogP based on the XGBoost model, respectively. ML: Machine learning; XGBoost: extreme gradient boosting.

the ML models with the DFT-calculated values. Figure 8 and Supplementary Table 18 provide a comparison of the predicted values from the RF, GBDT, and XGBoost models with the DFT-calculated values.

As shown in Figure 8, the ML models trained using the existing database demonstrate the ability to predict properties for unknown molecular datasets. Notably, the three ML models performed well in predicting the properties of the linear organic molecules TPA-TVT-TPA, YZ18, and YZ22. However, the predictions for spiro-OMeTAD, DTPC8-ThDTPA, DTPC13-ThTPA, and DTP-C6Th were less accurate, primarily due to the significant differences in the molecular structures of these molecules compared to those in the training set. The existing training dataset comprises only linear molecular structures, leading to a lack of diversity in the data. This limitation restricts the model's generalizability and its ability to understand and predict the properties of nonlinear molecular structures. Future efforts could focus on enhancing the diversity of the

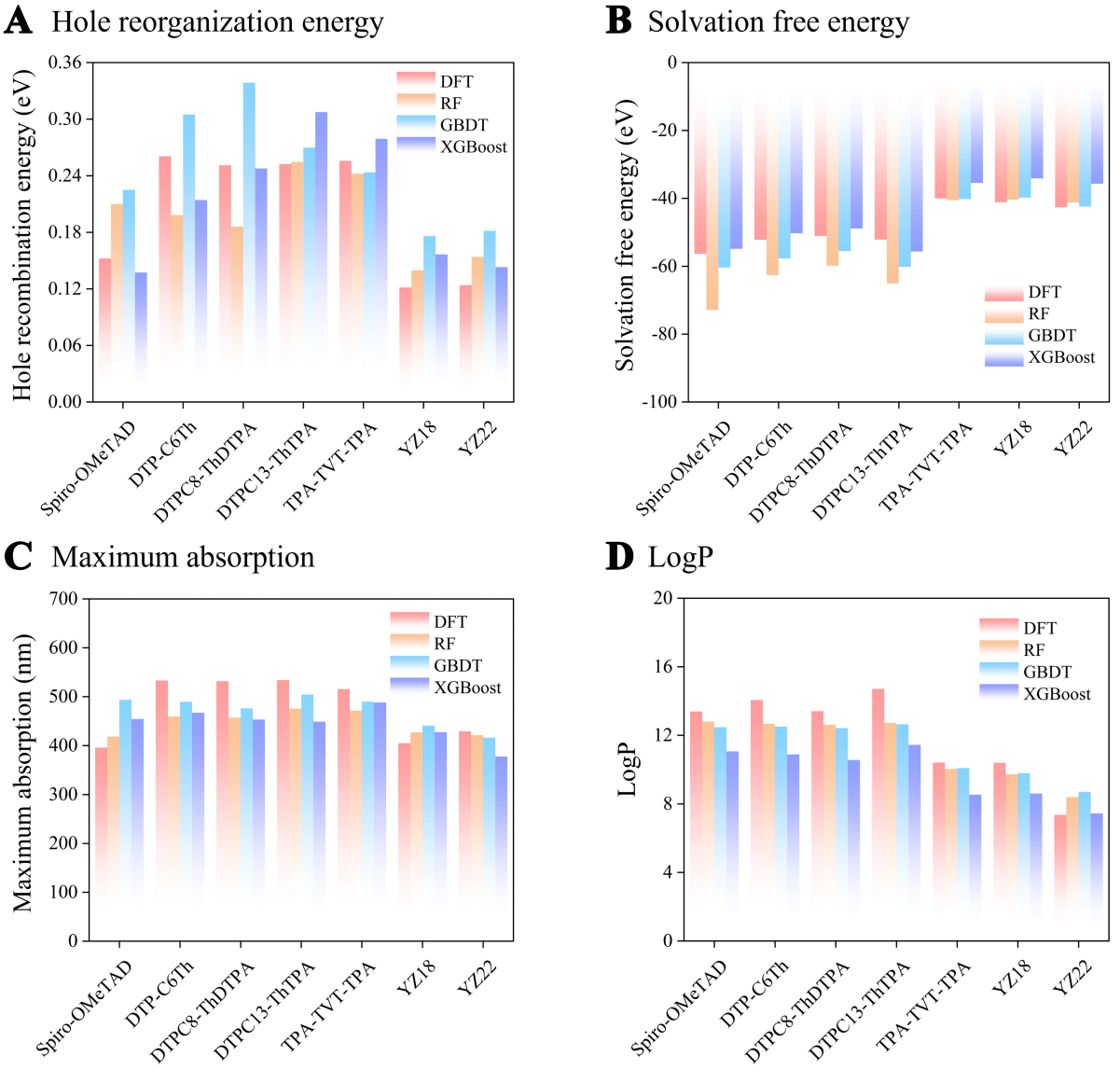

Figure 8. Comparison of the calculated DFT values for (A) Hole reorganization energy, (B) Solvation free energy, (C) Maximum absorption, and (D) LogP of common HTMs with the predicted values from RF, GBDT and XGBoost ML models, respectively. DFT: Density functional theory; HTMs: hole transport materials; RF: random forest; GBDT: gradient boosted decision tree; XGBoost: extreme gradient boosting; ML: machine learning.

training dataset by incorporating samples with distinct geometric structures (e.g., helical, star-shaped) to improve the model's capacity for learning across a wide range of molecular configurations. Additionally, applying structural transformation simulations to the existing data could generate diverse molecular datasets, further enriching the distribution of the training dataset. Alternatively, employing universal molecular descriptors or advanced feature extraction methods (such as neural networks or ensemble methods) could enable the model to better capture the fundamental characteristics of various molecular structures. Neural networks offer significant advantages in handling nonlinear molecular structures and complex features, enabling better capture of intricate relationships between molecules and thereby improving prediction accuracy. Ensemble methods, on the other hand, enhance prediction robustness by combining the outputs of multiple models, effectively reducing model bias and increasing the stability and

reliability of property predictions for SM-HTMs. These strategies aim to improve the model's generalization capabilities across different molecular architectures and enhance the accuracy of its performance predictions. On the other hand, as shown in Supplementary Table 18, the RF model and XGBoost model exhibited the best performance in predicting hydrophobicity and hole reorganization energy, respectively. On the other hand, the GBDT model performed better in predicting solvation free energy and maximum light absorption peak. Overall, the RF, GBDT, and XGBoost models, trained on the database obtained from high-throughput calculations, demonstrate good generalizability in predicting the material properties of linear organic SM-HTMs.

Methods for the design and development of new SM-HTMs

The discussion above, exemplified by linear SM-HTMs, presents a novel strategy for the design and development of SM-HTMs. The corresponding workflow is schematically illustrated in Figure 9. This methodology outlines a systematic and iterative approach designed to expedite the discovery of high-performance SM-HTMs. The process begins with the construction of a diverse library of candidate organic small molecules using a MSA. High-throughput computational methods are then applied to comprehensively evaluate the performance parameters of these molecules. At this stage, high-performing candidates can be screened and selected for further study. Additionally, the computational data generated can be used to train ML models, enabling the development of robust structure-property relationship models. This facilitates the direct prediction of performance parameters based on molecular structures generated by the splicing algorithm, significantly reducing dependence on resource-intensive simulations. If the model encounters new molecular structures that lead to inaccuracies in prediction, high-throughput computational methods can be reintroduced to optimize and refine the ML model. Moreover, the ML model can be employed for inverse design, allowing the identification of molecular structures predicted to exhibit superior performance. This integrated strategy effectively combines the computational efficiency of ML with the precision of high-throughput calculations, fostering iterative improvement in both predictive accuracy and material discovery. This seamless and adaptive workflow significantly accelerates the identification, screening, and optimization of next-generation SM-HTMs. While this study primarily focuses on the application of the proposed methodology to the design and screening of SM-HTMs for PSCs, the approach can be extended to other photovoltaic materials and device architectures. The MSA, combined with high-throughput computational screening and ML, provides a versatile framework that can be adapted for the discovery of new functional materials in various optoelectronic applications. For instance, in organic photovoltaics (OPVs), donor-acceptor molecular systems play a crucial role in determining device performance. By applying the MSA, high-throughput computational screening and ML strategies, a diverse library of donor-acceptor molecules can be systematically generated and screened based on key parameters such as frontier molecular orbital energies, exciton binding energy, and charge transport properties. Similarly, in dye-sensitized solar cells (DSSCs), this methodology can facilitate the identification of novel organic dyes with enhanced light absorption, redox stability, and efficient charge transfer properties. It offers a powerful framework for advancing material innovation in applications such as PSCs and other cutting-edge photovoltaic technologies.

CONCLUSIONS

In summary, this work presents a novel design strategy for key HTMs in PSCs, utilizing the combination of molecular splicing, high-throughput computational screening, and ML techniques to identify candidate molecular materials with outstanding structures, comprehensive properties, and synthetic feasibility. Approximately 200,000 π-type molecular structures were generated using a MSA, from which 7,399 molecules were selected for D-π-D-type molecular construction, followed by high-precision DFT calculations. Ultimately, a database of 7,222 D-π-D HTMs was compiled, containing property data for molecular structure models, HOMO levels, hole reorganization energy, solvation free energy, maximum

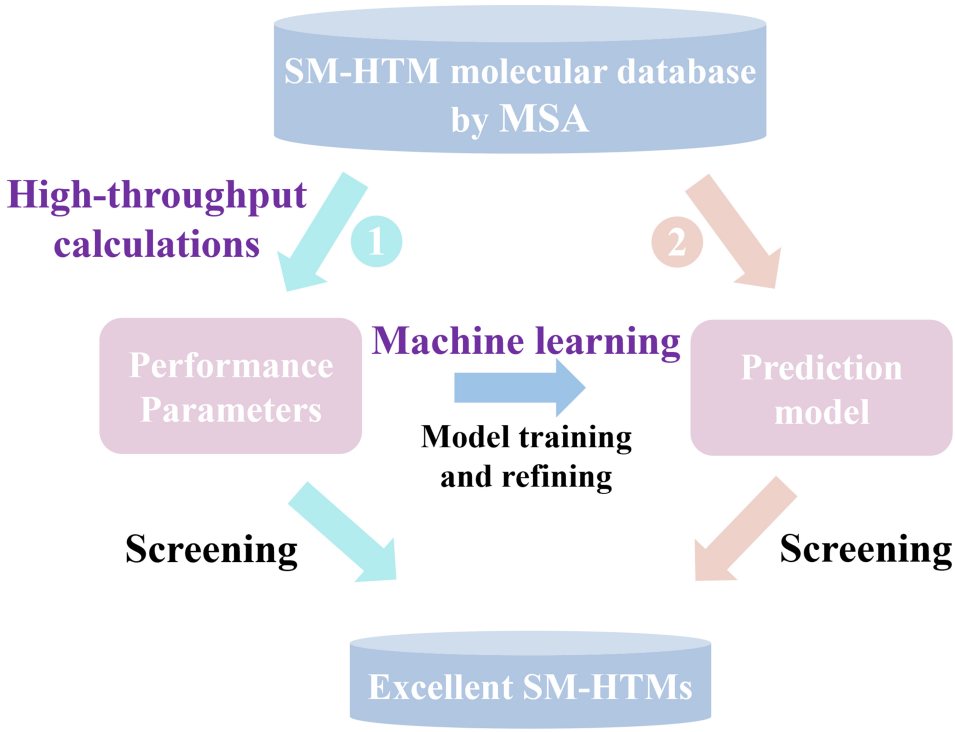

Figure 9. The toolkit for the design and development of novel SM-HTMs developed in this study. SM-HTMs: Small-molecule hole transport materials.

light absorption peak, hydrophobicity, and synthetic feasibility. Based on these property parameters, 6 molecules with excellent overall properties were selected for further synthesis and investigation. Additionally, RF, GBDT, and XGBoost ML models were developed using the molecular datasets created in this study. Among these, the XGBoost model demonstrated superior generalization ability and efficiency, achieving R^2 values of 0.901 for hole reorganization energy, 0.998 for solvation free energy, 0.969 for maximum light absorption peak, and 0.996 for hydrophobicity. Furthermore, the ML models trained in this work exhibited strong predictive performance for linear organic SM-HTMs similar to those in the dataset. This study provides a universal methodology for designing and developing SM-HTMs, which will fulfill the urgent demand for accelerating the progress of PSCs.

DECLARATIONS

Authors' contributions

Investigation, formal analysis, writing - original draft: Wen, J.
Conceptualization, data curation, validation: Yang, S.
Supervision, visualization, writing - review and editing: Jiang, L.
Data curation, investigation: Shi, Y.
Validation: Huang, Z.
Supervision: Li, P.; Xiong, H.
Validation, supervision: Yu, Z.
Validation, methodology, resources: Zhao, X.
Validation, funding acquisition, software: Xu, B.
Validation, writing - review and editing, funding acquisition, software, project administration: Wu, B.

Visualization, funding acquisition, project administration, software, writing - review and editing: Sa, B. Supervision, writing - review and editing: Qiu, Y.

Availability of data and materials

All datasets generated for this study are included in the article/Supplementary Materials.

Financial support and sponsorship

This work is financially supported by the National Key Research and Development Program of China (No. 2022YFB3807200), the National Natural Science Foundation of China (No. 51301039), the State Administration for Market Regulation (No. 2021MK050), and Fujian Provincial Department of Science & Technology of China (Nos. 2024J01948, 2023H6037, 2023S0065, 2021H6011).

Conflicts of interest

Yang, S.; Huang, Z.; Yu, Z.; Zhao, X. and Xu, B. are affiliated with Contemporary Amperex Technology Co., Limited (CATL). Sa, B. is the Guest Editor of the special issue "Advances in Machine Learning for Photoelectric Materials Research and Applications" in *Journal of Materials Informatics*. Sa, B. was not involved in any steps of the editorial process, including reviewer selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Bellani, S.; Bartolotta, A.; Agresti, A.; et al. Solution-processed two-dimensional materials for next-generation photovoltaics. *Chem. Soc. Rev.* **2021**, *50*, 11870-965. DOI PubMed PMC
2. Liu, Y.; Li, Y.; Wu, Y.; et al. High-efficiency silicon heterojunction solar cells: materials, devices and applications. *Mater. Sci. Eng. R. Rep.* **2020**, *142*, 100579. DOI
3. Lu, T.; Li, M.; Lu, W.; Zhang, T. Recent progress in the data-driven discovery of novel photovoltaic materials. *J. Mater. Inf.* **2022**, *2*, 7. DOI
4. Younis, A.; Lin, C. H.; Guan, X.; et al. Halide perovskites: a new era of solution-processed electronics. *Adv. Mater.* **2021**, *33*, 2005000. DOI
5. Huang, Y.; Liu, T.; Li, D.; et al. Limitations and solutions for achieving high-performance perovskite tandem photovoltaics. *Nano. Energy.* **2021**, *88*, 106219. DOI
6. Kim, Y. H.; Park, J.; Kim, S.; et al. Exploiting the full advantages of colloidal perovskite nanocrystals for large-area efficient light-emitting diodes. *Nat. Nanotechnol.* **2022**, *17*, 590-7. DOI
7. Wen, J.; Rong, K.; Jiang, L.; et al. Copper-based perovskites and perovskite-like halides: a review from the perspective of molecular level. *Nano. Energy.* **2024**, *128*, 109802. DOI
8. Green, M. A.; Dunlop, E. D.; Yoshita, M.; et al. Solar cell efficiency tables (Version 64). *Prog. Photovolt.* **2024**, *32*, 425-41. DOI
9. Zhang, C.; Wei, K.; Hu, J.; et al. A review on organic hole transport materials for perovskite solar cells: structure, composition and reliability. *Mater. Today.* **2023**, *67*, 518-47. DOI
10. Kim, G.; Choi, H.; Kim, M.; Lee, J.; Son, S. Y.; Park, T. Hole transport materials in conventional structural (n–i–p) perovskite solar cells: from past to the future. *Adv. Energy. Mater.* **2020**, *10*, 1903403. DOI
11. Sun, Q.; Sadhu, A.; Lie, S.; Wong, L. H. Critical review of Cu-based hole transport materials for perovskite solar cells: from theoretical insights to experimental validation. *Adv. Mater.* **2024**, *36*, e2402412. DOI PubMed
12. Li, M.; Qiu, F.; Wang, S.; Jiang, Y.; Hu, J. Hole transporting materials in inorganic $CsPbI_{3-x}Br_x$ solar cells: fundamentals, criteria and opportunities. *Mater. Today.* **2022**, *52*, 250-68. DOI

13. Yin, X.; Song, Z.; Li, Z.; Tang, W. Toward ideal hole transport materials: a review on recent progress in dopant-free hole transport materials for fabricating efficient and stable perovskite solar cells. *Energy. Environ. Sci.* **2020**, *13*, 4057-86. DOI
14. Kim, H. J.; Phenrat, T.; Tilton, R. D.; Lowry, G. V. Effect of kaolinite, silica fines and pH on transport of polymer-modified zero valent iron nano-particles in heterogeneous porous media. *J. Colloid. Interface. Sci.* **2012**, *370*, 1-10. DOI PubMed
15. Chang, Q.; Yun, Y.; Cao, K.; et al. Highly efficient and stable perovskite solar modules based on $FcPF_6$ engineered spiro-OMeTAD hole transporting layer. *Adv. Mater.* **2024**, *36*, 2406296. DOI
16. Dong, Y.; Rombach, F. M.; Min, G.; et al. Dopant-induced interactions in spiro-OMeTAD: advancing hole transport for perovskite solar cells. *Mater. Sci. Eng. R. Rep.* **2025**, *162*, 100875. DOI
17. Yang, H.; Shen, Y.; Zhang, R.; et al. Composition-conditioning agent for doped spiro-OMeTAD to realize highly efficient and stable perovskite solar cells. *Adv. Energy. Mater.* **2022**, *12*, 2202207. DOI
18. Rombach, F. M.; Haque, S. A.; Macdonald, T. J. Lessons learned from spiro-OMeTAD and PTAA in perovskite solar cells. *Energy. Environ. Sci.* **2021**, *14*, 5161-90. DOI
19. Khan, D.; Liu, X.; Qu, G.; Nath, A. R.; Xie, P.; Xu, Z. X. Nexuses between the chemical design and performance of small molecule dopant-free hole transporting materials in perovskite solar cells. *Small* **2023**, *19*, e2205926. DOI PubMed
20. Guo, H.; Zhang, H.; Shen, C.; et al. A coplanar π-extended quinoxaline based hole-transporting material enabling over 21 % efficiency for dopant-free perovskite solar cells. *Angew. Chem. Int. Ed. Engl.* **2021**, *60*, 2674-9. DOI
21. Yang, K.; Liao, Q.; Huang, J.; et al. Intramolecular noncovalent interaction-enabled dopant-free hole-transporting materials for high-performance inverted perovskite solar cells. *Angew. Chem. Int. Ed. Engl.* **2022**, *61*, 202113749. DOI
22. Yu, X.; Gao, D.; Li, Z.; et al. Green-solvent processable dopant-free hole transporting materials for inverted perovskite solar cells. *Angew. Chem. Int. Ed. Engl.* **2023**, *62*, 202218752. DOI
23. Jeong, M.; Choi, I. W.; Yim, K.; et al. Large-area perovskite solar cells employing spiro-Naph hole transport material. *Nat. Photon.* **2022**, *16*, 119-25. DOI
24. Yu, W.; Yang, Q.; Zhang, J.; et al. Simple is best: a *p*-phenylene bridging methoxydiphenylamine-substituted carbazole hole transporter for high-performance perovskite solar cells. *ACS. Appl. Mater. Interfaces.* **2019**, *11*, 30065-71. DOI
25. Qiu, J.; Liu, H.; Li, X.; Wang, S.; Zhang, F. Impact of 9-(4-methoxyphenyl) carbazole and benzodithiophene cores on performance and stability for perovskite solar cells based on dopant-free hole-transporting materials. *Solar. RRL.* **2019**, *3*, 1900202. DOI
26. Liu, X.; Ding, X.; Ren, Y.; et al. A star-shaped carbazole-based hole-transporting material with triphenylamine side arms for perovskite solar cells. *J. Mater. Chem. C.* **2018**, *6*, 12912-8. DOI
27. Zhang, F.; Liu, X.; Yi, C.; et al. Dopant-free donor (D)-π-D-π-D conjugated hole-transport materials for efficient and stable perovskite solar cells. *ChemSusChem* **2016**, *9*, 2578-85. DOI
28. Zhang, R.; Rong, F.; Lai, G.; Wu, G.; Ye, Y.; Zheng, J. Machine learning descriptors for crystal materials: applications in Ni-rich layered cathode and lithium anode materials for high-energy-density lithium batteries. *J. Mater. Inf.* **2024**, *4*, 17. DOI
29. Shi, Y.; Zhang, Y.; Wen, J.; et al. Interpretable machine learning for stability and electronic structure prediction of Janus III–VI van der Waals heterostructures. *Mater. Genome. Eng. Adv.* **2024**, *2*, e76. DOI
30. Shang, Y.; Xiong, Z.; An, K.; Hauch, J. A.; Brabec, C. J.; Li, N. Materials genome engineering accelerates the research and development of organic and perovskite photovoltaics. *Maters. Genome. Eng. Adv.* **2024**, *2*, e28. DOI
31. Gan, Y.; Miao, N.; Lan, P.; Zhou, J.; Elliott, S. R.; Sun, Z. Robust design of high-performance optoelectronic chalcogenide crystals from high-throughput computation. *J. Am. Chem. Soc.* **2022**, *144*, 5878-86. DOI PubMed
32. Sa, B.; Hu, R.; Zheng, Z.; et al. High-throughput computational screening and machine learning modeling of Janus 2D III–VI van der Waals heterostructures for solar energy applications. *Chem. Mater.* **2022**, *34*, 6687-701. DOI
33. Xu, J.; Chen, H.; Grater, L.; et al. Anion optimization for bifunctional surface passivation in perovskite solar cells. *Nat. Mater.* **2023**, *22*, 1507-14. DOI
34. Zhang, B.; Zeng, H.; Yin, H.; et al. Combining component screening, machine learning and molecular engineering for the design of high-performance inverted perovskite solar cells. *Energy. Environ. Sci.* **2024**, *17*, 5532-41. DOI
35. Wang, X.; Wang, M.; Zhang, Z.; et al. De novo design of spiro-type hole-transporting material: anisotropic regulation toward efficient and stable perovskite solar cells. *Research* **2024**, *7*, 0332. DOI PubMed PMC
36. Sun, Z.; Long, R. Thia[5]helicene-based D−π–A-type molecular semiconductors for stable and efficient perovskite solar cells: a theoretical study. *J. Phys. Chem. C.* **2023**, *127*, 8953-62. DOI
37. Sun, Z. Z.; Li, Y.; Xu, X. L. Donor engineering of a benzothiadiazole-based D-A-D-type molecular semiconductor for perovskite solar cells: a theoretical study. *Phys. Chem. Chem. Phys.* **2024**, *26*, 6817-25. DOI
38. Zhu, W.; Zhou, K.; Fo, Y.; et al. Rational design of small molecule hole-transporting materials with a linear π-bridge for highly efficient perovskite solar cells. *Phys. Chem. Chem. Phys.* **2022**, *24*, 18793-804. DOI
39. Paramasivam, G.; Sambasivam, S.; Kumar Ravva, M. Designing donor-acceptor-donor (D-A-D) type molecules for efficient hole-transporting in perovskite solar cells - a DFT study. *ChemistrySelect* **2023**, *8*, e202204462. DOI
40. Wu, J.; Torresi, L.; Hu, M.; et al. Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science* **2024**, *386*, 1256-64. DOI
41. Faruque, M. O.; Akter, S.; Limbu, D. K.; Kilway, K. V.; Peng, Z.; Momeni, M. R. High-throughput screening, crystal structure prediction, and carrier mobility calculations of organic molecular semiconductors as hole transport layer materials in perovskite solar cells. *Cryst. Growth. Des.* **2024**, *24*, 8950-60. DOI

42. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; et al. Gaussian 16 Rev. C.01[CP/OL]. https://gaussian.com/gaussian16/. (accessed on 10 Apr 2025)
43. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902-9. DOI
44. Raghavachari, K. Perspective on "Density functional thermochemistry. III. The role of exact exchange". *Theor. Chem. Acc.* **2000**, *103*, 361-3. DOI
45. Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Account.* **2008**, *120*, 215-41. DOI
46. Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. Assessment of the performance of the M05-2X and M06-2X exchange-correlation functionals for noncovalent interactions in biomolecules. *J. Chem. Theory. Comput.* **2008**, *4*, 1996-2000. DOI PubMed
47. Marcus, R. A. Electron transfer reactions in chemistry: theory and experiment (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 1111-21. DOI
48. Hutchison, G. R.; Ratner, M. A.; Marks, T. J. Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects. *J. Am. Chem. Soc.* **2005**, *127*, 2339-50. DOI PubMed
49. Marcus, R.; Sutin, N. Electron transfers in chemistry and biology. *Biochim. Biophys. Acta. Rev. Bioenerg.* **1985**, *811*, 265-322. DOI
50. Ho, J.; Klamt, A.; Coote, M. L. Comment on the correct use of continuum solvent models. *J. Phys. Chem. A.* **2010**, *114*, 13442-4. DOI PubMed
51. Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016-27. DOI
52. Liu, B.; Jin, J.; Liu, M. Mapping structure-property relationships in fullerene systems: a computational study from C20 to C60. *npj. Comput. Mater.* **2024**, *10*, 1410. DOI
53. Bannan, C. C.; Calabró, G.; Kyu, D. Y.; Mobley, D. L. Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. *J. Chem. Theory. Comput.* **2016**, *12*, 4015-24. DOI PubMed PMC
54. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8. DOI PubMed PMC
55. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: machine learning in python. *arXiv* **2012**, arXiv:1201.0490. Available online: https://doi.org/10.48550/arXiv.1201.0490. [accessed 10 Apr 2025]
56. Chen, T.; Guestrin, C. XGBoost: a scalable tree boosting system. *arXiv* **2012**, arXiv:1603.02754. Available online: https://doi.org/10.48550/arXiv.1603.02754. [accessed 10 Apr 2025]
57. Pan, T.; Li, Z.; Ren, B.; et al. Stabilizing doped Spiro-OMeTAD with an organic molten salt for efficient and stable perovskite solar cells. *Energy. Environ. Sci.* **2024**, *17*, 9548-54. DOI
58. Ren, Y.; Wei, Y.; Li, T.; et al. Spirobifluorene with an asymmetric fluorenylcarbazolamine electron-donor as the hole transport material increases thermostability and efficiency of perovskite solar cells. *Energy. Environ. Sci.* **2023**, *16*, 3534-42. DOI
59. Zhang, T.; Wang, F.; Kim, H. B.; et al. Ion-modulated radical doping of spiro-OMeTAD for more efficient and stable perovskite solar cells. *Science* **2022**, *377*, 495-501. DOI
60. Ren, M.; Fang, L.; Zhang, Y.; et al. Durable perovskite solar cells with 24.5% average efficiency: the role of rigid conjugated core in molecular semiconductors. *Adv. Mater.* **2024**, *36*, e2403403. DOI
61. Dong, Z.; Yin, X.; Ali, A.; et al. A dithieno[3,2-*b*:2',3'-*d*]pyrrole-cored four-arm hole transporting material for over 19% efficiency dopant-free perovskite solar cells. *J. Mater. Chem. C.* **2019**, *7*, 9455-9. DOI
62. Zhou, J.; Yin, X.; Dong, Z.; et al. Dithieno[3,2-b:2',3'-d]pyrrole cored p-type semiconductors enabling 20% efficiency dopant-free perovskite solar cells. *Angew. Chem. Int. Ed. Engl.* **2019**, *58*, 13717-21. DOI
63. Yin, X.; Zhou, J.; Song, Z.; et al. Dithieno[3,2-b:2',3'-d]pyrrol-cored hole transport material enabling over 21% efficiency dopant-free perovskite solar cells. *Adv. Funct. Mater.* **2019**, *29*, 1904300. DOI
64. Zhao, B. X.; Yao, C.; Gu, K.; Liu, T.; Xia, Y.; Loo, Y. A hole-transport material that also passivates perovskite surface defects for solar cells with improved efficiency and stability. *Energy. Environ. Sci.* **2020**, *13*, 4334-43. DOI
65. Pham, H. D.; Hu, H.; Feron, K.; et al. Thienylvinylenethienyl and naphthalene core substituted with triphenylamines - highly efficient hole transporting materials and their comparative study for inverted perovskite solar cells. *Sol. RRL.* **2017**, *1*, 1700105. DOI
66. Xu, J.; Liang, L.; Mai, C. L.; et al. Lewis-base containing spiro type hole transporting materials for high-performance perovskite solar cells with efficiency approaching 20. *Nanoscale* **2020**, *12*, 13157-64. DOI
67. Sandoval-Torrientes, R.; Zimmermann, I.; Calbo, J.; et al. Hole transporting materials based on benzodithiophene and dithienopyrrole cores for efficient perovskite solar cells. *J. Mater. Chem. A.* **2018**, *6*, 5944-51. DOI