Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells

Jianchang Wu[1,2]*†, Luca Torresi[3,4]†, ManMan Hu[5]†, Patrick Reiser[3,4]†, Jiyun Zhang[1,2], Juan S. Rocha-Ortiz[1,2], Luyao Wang[6]*, Zhiqiang Xie[2], Kaicheng Zhang[2], Byung-wook Park[5], Anastasia Barabash[1,2], Yicheng Zhao[1,2,7], Junsheng Luo[2,7], Yunuo Wang[2], Larry Lüer[1,2], Lin-Long Deng[6], Jens A. Hauch[1,2], Dirk M. Guldi[8], M. Eugenia Pérez-Ojeda[9], Sang Il Seok[5]*, Pascal Friederich[3,4]*, Christoph J. Brabec[1,2,10]*

The inverse design of tailored organic molecules for specific optoelectronic devices of high complexity holds an enormous potential but has not yet been realized. Current models rely on large data sets that generally do not exist for specialized research fields. We demonstrate a closed-loop workflow that combines high-throughput synthesis of organic semiconductors to create large datasets and Bayesian optimization to discover new hole-transporting materials with tailored properties for solar cell applications. The predictive models were based on molecular descriptors that allowed us to link the structure of these materials to their performance. A series of high-performance molecules were identified from minimal suggestions and achieved up to 26.2% (certified 25.9%) power conversion efficiency in perovskite solar cells.

The design of hole-transporting materials (HTMs) for perovskite solar cells (PSCs) has mainly been driven by experimentalists qualitatively recognizing patterns in HTM structures to improve device performance (*1*–*3*). This approach lacks a mechanistic understanding of new HTMs but also requires pattern recognition in high-dimensional datasets. Machine learning (ML) has been used to detect meaningful patterns for various applications in science and technology, including organic synthesis (*4*, *5*), materials science (*6*–*9*), and fabrication process optimization (*10*, *11*). However, the discovery of new materials with optimized properties for semiconducting device functionality (*11*–*13*) has not been applied to emerging photovoltaics. Inverting the relation between device performance and a material's structure is difficult because of the complex correlation between the material's structural features, the processing-dominated microstructure of composites, and the relative impact of both on device performance.

Prior efforts have focused primarily on the use of ML to optimize the fabrication process or to predict device performance and stability on the basis of fabrication processes. For example, Gaussian process (GP) regression has been used to model data from robotic device fabrication, which has enabled the analysis and prediction of device performance and stability (*14*–*17*). The best parameter set and objective function could be quickly identified over the entire parameter space with a minimal number of samples. A similar approach was applied by Xu *et al.* to optimize the passivation materials for perovskite (*18*). However, the training data are limited to the fabrication process or commercial materials and, as a result, did not include the generation of new molecular structures.

Two recent studies combined ML and organic synthesis. Bai *et al.* reported a successful case in which a gradient-boosted tree regressor model, trained on data from 170 synthesized conjugated polymers, accurately predicted high-performance photocatalysts from a virtual database of 6354 candidates (*19*). The trend obtained from the training dataset was verified by experimental data on the newly synthesized polymers, leading to further refinement of the model. However, the insoluble characteristics of those polymers, although reducing the challenges of purification and enriching the database, have limited the broader applications of this material class. Gómez-Bombarelli *et al.* integrated virtual screening, cheminformatics, and organic synthesis to predict emitters for organic light–emitting diodes (*12*). At that time, the overall data volume and iteration number within the closed loop were restricted by the number of synthesized molecules.

One of the general findings to emerge from these most recent studies is that the autonomous optimization algorithms require not only sufficiently large data volume but also data diversity, which necessitates the possibility of synthesizing structurally diverse molecules. Given the multidimensional nature of optimizing a chemical structure for device performance, the challenge is generating sufficiently large and consistent datasets to ensure the implementation of these algorithms (*9*, *20*, *21*).

To tackle these problems, we have developed a joint knowledge- and data-based strategy, implemented in a high-throughput (HT) organic synthesis platform, that can synthesize and purify >100 solution-processable small-molecule semiconductors with varying structures and consistent quality over multiple synthesis campaigns within a time frame of weeks (*22*). In particular, the minor fluctuations of <3% observed in PSCs with a power conversion efficiency (PCE) of 21% provide a solid foundation for the smooth operation of the entire workflow. The knowledge-based part is pivotal to design the HT synthesis platform, clarify purification and characterization of the molecules, and implement all the necessary analytics to record properties expected to impact device performance. This approach provided us with a sufficient large and diverse dataset for implementing the data-based exploration strategy by training an ML model coupling the structural features of the HTM to the performance of corresponding p-i-n PSCs. We developed a workflow that coupled automatic HT experiments, ML, and validation through further HT experiments. These feedback loops predicted new structures based on device results and allowed for a closed optimization of the material to the target criteria of the solar cell. We could rapidly acquire sufficient quantities of new organic conjugated molecules to enable multiple iterations of experimental data.

The model, trained by 149 synthesized molecules, accurately predicted high-performance HTMs from a virtual database of 1 million candidates. Given the already considerable complexity of this study, we did not undertake an individual optimization of the device structure for each new HTM during the Bayesian optimization (BO) operation. Instead, device optimization was done for a series of HTMs that showed an initial PCE exceeding 20% after BO operation. These materials finally reached PCEs of >26% (certified at 25.9%), highlighting the enormous potential of coupling material and device optimization in a combined workflow.

Workflow

We used Suzuki coupling to HT synthesize new molecules. This reaction facilitates the combination of two distinct monomers, A- and

[1]Forschungszentrum Jülich GmbH, Helmholtz-Institute Erlangen–Nürnberg (HI-ERN), Erlangen, Germany. [2]Faculty of Engineering, Department of Material Science, Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander-Universität Erlangen–Nürnberg (FAU), Erlangen, Germany. [3]Institute of Nanotechnology, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. [4]Institute of Theoretical Informatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. [5]Department of Energy Engineering, School of Energy and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea. [6]State Key Lab for Physical Chemistry of Solid Surfaces, Department of Chemistry, College of Chemistry and Chemical Engineering, Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen, China. [7]National Key Laboratory of Electronic Films and Integrated Devices, School of Integrated Circuit Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. [8]Department of Chemistry and Pharmacy & Interdisciplinary Center of Molecular Materials (ICMM), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. [9]Department of Chemistry and Pharmacy, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. [10]Zernike Institute for Advanced Materials, University of Groningen, Groningen, Netherlands.
***Corresponding author. Email: jianchang.wu@fau.de (J.W.); wangluyaoyz@gmail.com (L.W); seoksi@unist.ac.kr (S.I.S); pascal.friederich@kit.edu (P.F.); christoph.brabec@fau.de (C.J.B.)**
†These authors contributed equally to this work.

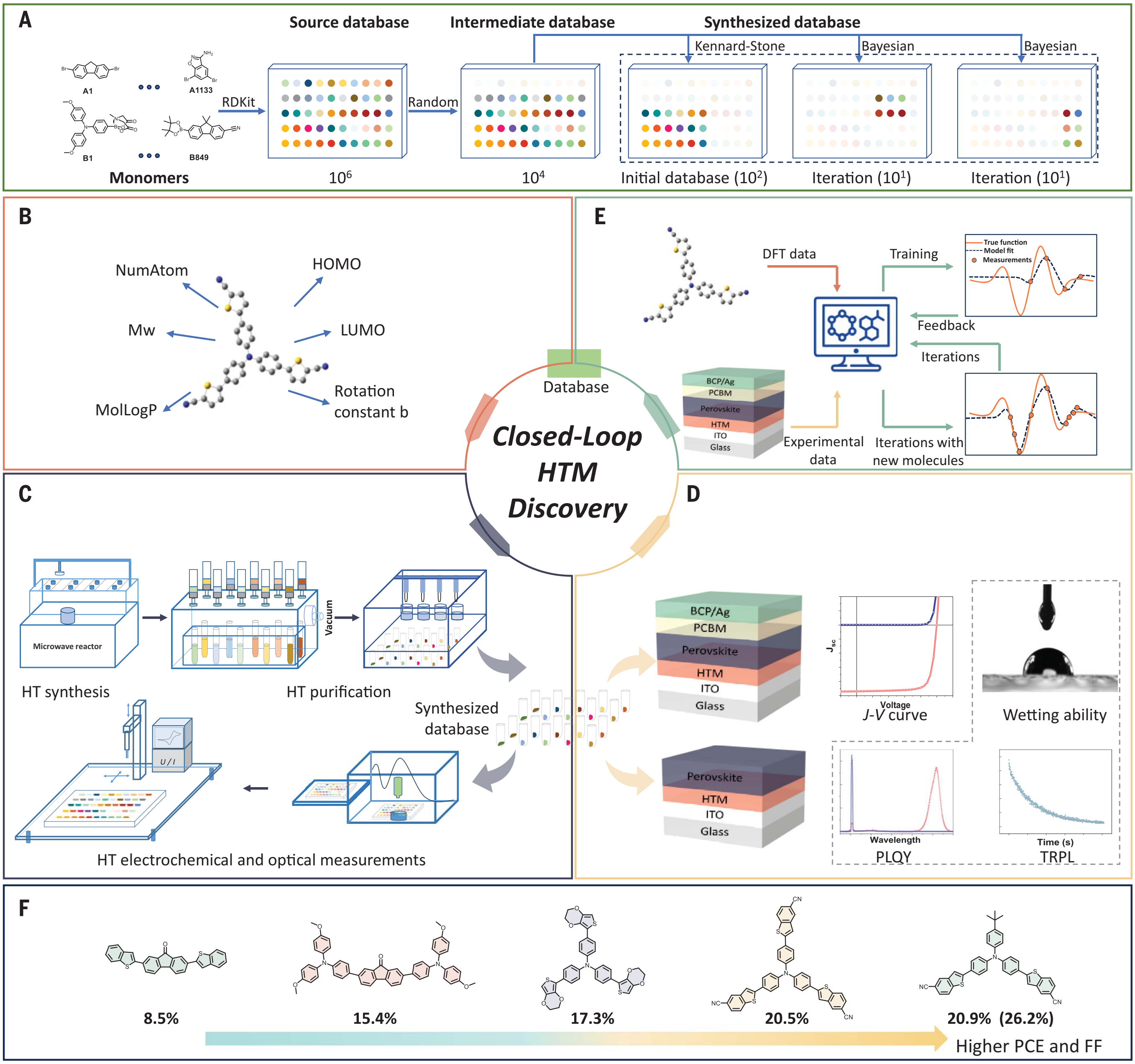
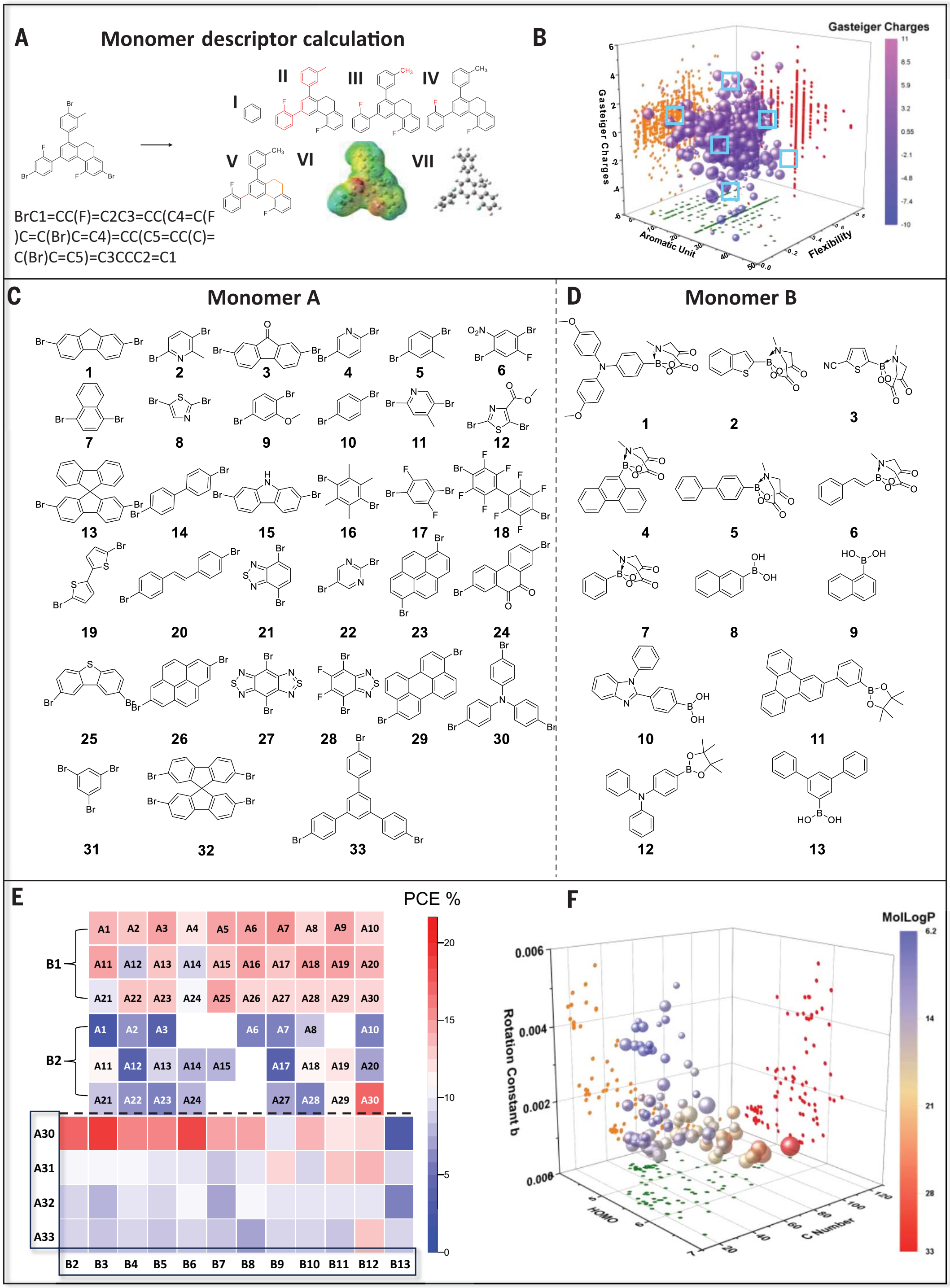


Fig. 1. Approach overview. (**A**) We used three kinds of databases. The source database was the virtual combination of two types of commercial monomers using the Suzuki coupling rule. The intermediate database contained randomly selected molecules from the source database for DFT calculations. The synthesized database included synthesized molecules used in this study, including an initial database for model training and two iteration databases for model validation and correction. (**B**) DFT calculations provided descriptors of molecules in the intermediate database. NumAtom, number of atoms; Mw, molecular weight; MolLogP, molecular logarithm of partition coefficient. (**C**) Molecules in the synthesized database were synthesized, purified, and characterized through our in house high throughput (HT) platform. (**D**) The synthesized molecules were used as HTMs in PSCs and characterized in devices and semidevices. ITO, indium tin oxide; BCP/Ag, bathocuproine and sliver. (**E**) The model was trained on HTM descriptors and device parameters. New molecules were predicted, synthesized and experimentally measured, and fed back to the database. The iteration was repeated until the discovery of the best HTM from the set. (**F**) Molecular iterations were summarized and analyzed.

B-type molecules, into a B-A-B–type conjugated molecule. The workflow (Fig. 1) began with the creation of a source database and the definition of subdatabases. The source database combined all commercially available monomers A and B compatible with Suzuki coupling of organic bromines with boronic acids. The intermediate database, containing results of density functional theory (DFT) calculations, consists of 13,000 randomly selected molecules from the source database. The synthetic database was then selected from the intermediate database according to specific rules implanted with a Kennard-Stone algorithm for the initial database and BO for iteration database.

The initial pretraining was executed for representative monomers (33 As and 13 Bs) that were selected to ensure the broadest coverage of chemical features. We then performed DFT calculations of robust chemical descriptors for

Fig. 2. Generation of the initial library. (**A**) Monomer descriptor calculation from different perspectives. (I) Aromatic ring species; (II) conjugate length; (III) substituent species; (IV) active group; (V) flexible and rigid units; (VI) electronic effect; (VII) spatial effect. (**B**) Monomer subset selection (blue ball) with the Kennard Stone algorithm from a commercial monomer library (square). Selected A (**C**) and B (**D**) monomers for the initial database. (**E**) Color map of PCEs for HTMs in the initial library. It consists of two parts: the combination of A1 to A30 with B1 to B2 (top) and A30 to A33 with B2 to B13 (bottom). In the top portion, the serial number of each monomer A is filled in the cell. (**F**) In silico library for HTM descriptors used for initial library selection.


A
Monomer descriptor calculation
I
II
III
IV
V
VI
VII
BrC1=CC(F)=C2C3=CC(C4=C(F)C=C(Br)C=C4)=CC(C5=CC(C)=C(Br)C=C5)=C3CCC2=C1
B
Gasteiger Charges
Aromatic Unit
Flexibility
C
Monomer A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
D
Monomer B
1 2 3 4 5 6 7 8 9 10 11 12 13
E
PCE %
B1
B2
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10
A11 A12 A13 A14 A15 A16 A17 A18 A19 A20
A21 A22 A23 A24 A25 A26 A27 A28 A29 A30
A1 A2 A3 A6 A7 A8 A10
A11 A12 A13 A14 A15 A17 A18 A19 A20
A21 A22 A23 A24 A27 A28 A29 A30
A30
A31
A32
A33
B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13
20
15
10
5
0
F
MolLogP
Rotation Constant b
HOMO
C Number
0.006
0.004
0.002
0.000

each molecule. Next, high-throughput organic synthesis was executed in four campaigns along with purification and characterization to create an initial database of 101 molecules. We used these molecules to fabricate and characterize p-i-n perovskite devices and then used ML based on calculated descriptors and experimental data to predict the performance of each new molecule within the virtual compound library. The model was then validated by synthesizing 48 molecules suggested by the BO in two further campaigns from a space covering 50 A and 19 B molecules. This process was repeated to explore further molecules. Lastly, the set of structure-property relations that connected device performance to molecular properties was built and analyzed by multitask GP regressors (MTGPR), combining data from HTM films as well as HTM-perovskite interface characterization. On the basis of these rules, molecular design principles were established that led to the iterative refinement of the molecular structure.

Initial library generation: Synthesis and device fabrication

As the first step, the formulation of an in silico library encompassing 1 million HTMs utilizing the RDKit package (see section S3 of the supplementary materials for details) was described. To amass a substantial dataset, we set two

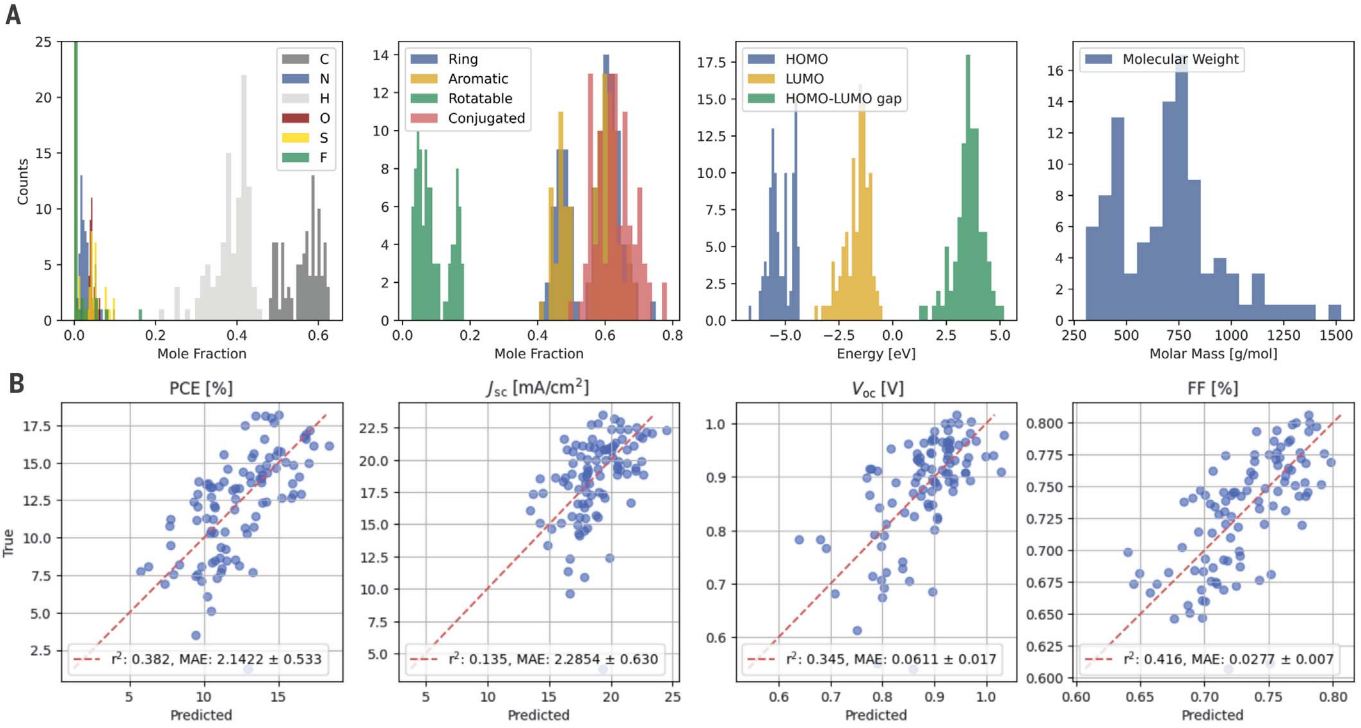


Fig. 3. Model training based on experimental data and in silico descriptors. (**A**) Distribution of important features from the dataset used in the model training. (**B**) The prediction accuracy of the GP for all device labels, including maximum PCE, J_{sc}, V_{oc}, and FF. MAE, mean absolute error.

primary criteria for monomer selection: (i) Reacting units (bromine for monomer A, boronic acid derivatives for monomer B) needed to be attached to aromatic moieties, and (ii) monomer B was restricted to only one reacting unit to prevent polymerization. No other selection criteria or any intuitive selection rules were applied.

Following these criteria, 1132 A and 850 B monomers were selected. Subsequently, a script based on RDKit was used to annotate the monomers across seven aspects, encompassing types of conjugated frameworks, substituent types, electronic effects, and steric effects (Fig. 2A). The categorization of HTMs along these seven aspects was aligned to the current state of understanding how HTMs impact perovskite device performance (*3*, *23*–*25*). A representative library of monomers (33 As and 13 Bs) was then chosen from this space by using the Kennard-Stone algorithm (Fig. 2, B to D) (*26*). This sampling method choice ensured that monomers are selected from uniform regions of the feature space. To guarantee comparability with the reported molecules, a few monomers with good reported performance were manually added.

To make more efficient use of the selected monomers and to compare the differences in their structure-performance correlation, we refrained from simple pairings. Instead, we divided the monomers into two groups for synthesis: A1 to A30 combined with B1 and B2 and A30 to A33 combined with B2 to B13. A comprehensive description of the synthesis, purification, and precharacterization was recently published (*22*). That routine was an essential pillar for the next step, the autonomous material discovery for optimized device performance, which we report in this work. We adopted a semiautomated synthesis platform in which a microwave reactor accelerated the synthetic process. Subsequently, the synthesized molecules underwent a two-step purification involving fast filtration and recrystallization.

To ensure immediate access to the identified promising candidates and assess the potential impact of synthesis and purification methods on material performance, the batch-to-batch reproducibility of the platform was verified by characterizing 10 representative molecules from different batches, including nuclear magnetic resonance (NMR), mass spectrometry, optoelectronic properties, and device performance (see section S2 in the supplementary materials). Minor fluctuations, with <3% for PCE and <1% for ultraviolet-visible (UV-Vis) absorption, were observed between the different batches, so the reproducibility of our experimental dataset did not limit the functionality of the ML algorithms. Such high reproducibility is further relevant for FAIR-guided data libraries (where FAIR is "findable, accessible, interoperable, and reusable") that would allow other scientists to reproduce our experiments.

These synthesized molecules were used as dopant-free HTMs in p-i-n–structured PSC devices, where the function of the HTM goes beyond merely extracting and transporting holes; it plays a crucial role in influencing the crystal growth of the perovskite (*23*, *27*, *28*). Perovskite absorber layers were grown from either $(CsI)_{0.05}(MAPbBr_3)_{0.17}(FAPbI_3)_{0.83}$ (MA, methylammonium; FA, formamidinium) or $(CsPbI_3)_{0.17}(FAPbI_3)_{0.83}$. We used [6,6]-phenyl-$C_{61}$-butyric acid methyl ester (PCBM) as the electron-transporting layer and silver as the top electrode. The possibility of HTM damage during the solution deposition of perovskite was carefully considered and evaluated before conducting device fabrication. Although some HTMs are soluble in N,N′-dimethylformamide (DMF)–dimethyl sulfoxide (DMSO) mixture, only a small amount was washed away during spin coating the pristine solvents (fig. S8). This is attributed to the short interaction time during spin coating but also to the relatively low solubility in DMF-DMSO mixture.

The resulting current density–voltage (*JV*) curves are depicted in fig. S11. Reference devices based on poly(triarylamine) (PTAA), a state-of-the-art commercial HTM for p-i-n PSCs, were used to calibrate the variations within each batch of devices. Furthermore, 6 to 12 devices

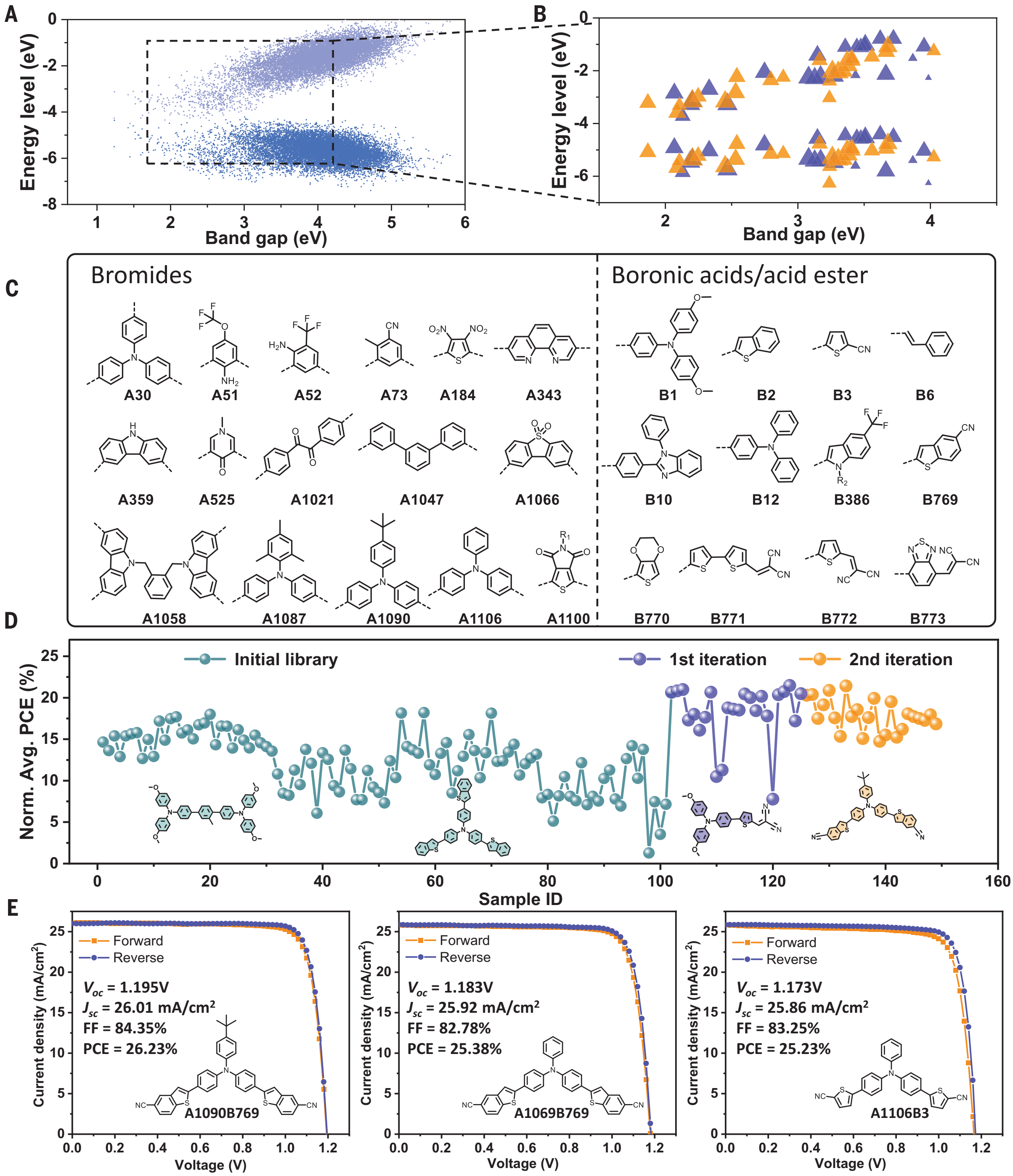


Fig. 4. New synthesized molecules and experimental data for iteration. (**A**) Predicted properties of the molecular library. (**B**) Selected molecules from the database for iterations based on calculated properties, where marker size is proportional to device performance. Blue triangles represent the molecules in the first iteration, and the orange ones, those in the second. (**C**) Molecular fragments used of iterative molecules. To visualize the structure of the final molecules, molecular fragments are used to represent the monomers. (**D**) Device performance of molecules for the initial dataset and the iterations. The PCE of the samples is the average performance normalized to the reference devices (based on PTAA) of the corresponding batch. (**E**) Further personalized optimization of the devices based on representative molecules. Blue, standardized condition; orange, personalized condition based on molecular properties.

were prepared per molecule to provide sufficient statistics on the data relevance. The average value was used as the experimental data point. To further ensure the validity of the data, reference devices were extensively optimized to minimize inter batch differences. The device performance parameters of each new molecule were normalized to the reference device with a PTAA hole transport layer before entering them into the ML model. Additionally, throughout the iterative process, the normalization to PTAA allowed us to adapt the perovskite recipe throughout the experimental campaign to achieve better reproducibility at the highest performance.

The heatmap of the PCEs for all synthesized molecules in Fig. 2E revealed the initial trends of interest. The PCEs of molecules A reacting with B1 were generally higher than A reacting with B2. Similarly, while keeping the monomer B constant, almost all molecules based on A30 as a central building block demonstrated the best performance. This result highlighted the substantial advantage of B1 and A30, triphenylamine (TPA) derivatives, in HTMs, but there are notable exceptions: A30B2 > A30B1, A31B12 > A30B12, and A31B13 > A30B13. The first and second points share a commonality: The monomers B1 and B12 have a TPA structure, which suggests that excess TPA or its placement on the periphery may be disadvantageous. As for the third point, it appears challenging to explain it from chemical intuition alone and highlights the need to use ML to provide further insights and identify underlying mechanisms.

ML models and feature engineering

To better understand structure-property relations in the observed data, we constructed an ML model that correlated representative molecular descriptors to the PCEs of the devices. In contrast to categorical labels, such as A and B, continuous molecular descriptors can be used to provide an ML-readable description that can integrate unseen A or B fragments into the same ontology. Typically, this process requires feature engineering to find meaningful descriptors for the problem. We generated a large virtual molecular library by evaluating the expected reaction products from a set of A and B building blocks with the RDKit software package. Subsequently, the three-dimensional geometry of the molecules was optimized by using a conformer search in CREST and the semiempirical density-functional tight-binding program xTB (*19*, *29*) followed by the calculation of molecular descriptors with DFT in TURBOMOLE (see section S3 in the supplementary materials for details).

For the ML model, we sought a set of descriptors that adequately captured device differences without relying on a specific hypothesis. To achieve this, we chose a combination of simple molecular statistics, such as the number of atom species, aromatic bonds, and specific functional groups; with theoretically computed features, such as the logarithm of solubility, molecular orbital energies, dipole moment; and geometric properties, such as rotational constants. A list of all features with explanations can be found in the supplementary materials.

In Fig. 3A, we show the distribution of the most important features. Although our descriptor acquisition was not based on a single specific hypothesis, several reported features were included in the descriptor set, mainly based on our understanding of how these features may impact the molecules properties. A fairly straight criterion was the DFT-calculated highest-occupied molecular orbital (HOMO) level, which is very well understood in terms of impacting device performance. An offset between the HOMO level of the HTL and the perovskite VB will lead to an extraction barrier, leading to a reduced hole extraction rate and, subsequently, to enhanced recombination at that interface. The influence on device performance is expected to follow a step-function or, rather, sigmoidal s-shape trend as a function of the energetic offset between the two electronic bands. Specific atoms, such as fluorine (F), and heterocycles, such as thiophene and aniline, were considered, as they were reported to have positive interactions with perovskites, such as a passivation effect (*30*). Additionally, factors affecting hole transport, such as molecular rigidity and conjugation, were considered. These features were expected to influence intermolecular interactions (*31*) and, therefore, were also included in our descriptor set.

For model selection, we trained different ML models on a random 10-fold cross-validation of the 101 experimental molecular data points. Tested ML models included random forest regression, linear regression, neural networks, GP regression, and kernel-ridge regression. All simple models performed equally well. For BO, we chose the GP as a surrogate model because it offered an uncertainty measure required in many acquisition strategies. The prediction accuracy of all device labels beyond PCE for the GP model is shown in Fig. 3B. A more detailed analysis of the influence of features and what physical insights can be learned from the ML models is discussed in “Model analysis.”

Experimental validation of the model

To demonstrate that the ML model can discover new molecules by predicting viable new organic semiconductors for hole extraction, we conducted two iterations of closed-loop materials optimization. This process entailed the identification of potential candidates through the ML surrogate model and Bayesian selection criteria, automatic synthesis of candidates, and, lastly, device characterization used to update the model. In the first iteration, 24 new molecules (blue triangles, Fig. 4A) were synthesized in a single batch and characterized to validate the previously obtained model. New building blocks (A525-pyridinone, A772-dicyanovinyl) and asymmetric structural motifs were considered to enrich the diversity of molecular structures in the database. The monomer database for synthesis was expanded from 33 × 13 to 50 × 19, mainly through ML recommendations, supplemented by accessibility, derivatives exhibiting high-performance structures in the initial database, and random acquisition.

The target for optimization was the device PCE averaged over 12 devices with an upper confidence bound (UCB) acquisition function, which should have identified high-performance HTMs for further exploration. We found that the new series of materials resulted in device PCEs that were generally higher than those of the materials in the initial database, demonstrating the advantage of ML over random sampling or grid search approaches when operated in the “exploit” mode. Among them, six molecules exceeded the PTAA-based device reference. We also observed outliers to the predictions (samples 110 and 120), which were separately examined (see section S4 in the supplementary materials) and could be explained in terms of impurities, unfavorable wetting behavior, or limited solubility.

For the second iteration, we increased the explorative character of the UCB in the model (*5*, *17*). Subsequently, molecules 126 to 149, recommended by ML, were synthesized (Fig. 4). Although no new champion HTM was found, the average PCE of predictions was still very high, comparable to that of the first iteration. This result affirmed the potential and viability of our workflow, considering the multitude of factors governing the performance of perovskite devices. Device performance is only partially limited by the HTM, as many other layers, including the electron-transporting layer, the electrodes, and the quality and defect density of the absorbing layer, can also reduce performance. Moreover, the physical models behind these losses include multiple chemical and physical phenomena, including inherently limiting material properties, generation, recombination and transport dynamics in semiconductors, the energetics of interface formation and the corresponding potential landscape in thin-film devices, thin-film microstructure formation, the device architecture itself, and the macroscopic film homogeneity. All of these aspects are described in the ontology of perovskite devices that has been recently uploaded on Matportal (*32*).

Given these aspects, we were surprised to see the wide spread in PCE among the various HTMs within our library. We expected a threshold-type behavior for HTM materials having a too-small bandgap or a too-small or even negative offset in the HOMO and lowest unoccupied molecular orbital (LUMO) potentials versus the valence band maximum and conduction band minimum of the perovskite. However, the broad variation of representative PCE values ranging from 15 to 21% postulates that the nature of the HTM molecules can determine the performance of perovskite devices beyond our expectations.

To obtain accurate trends, the performance of these molecules was characterized under uniform device conditions, with the standard device optimized based on PTAA. When exploring the limit of a material's performance, device parameters need to be adjusted according to the material's characteristics. Thus, we selected three promising molecules for individualized device performance optimization.

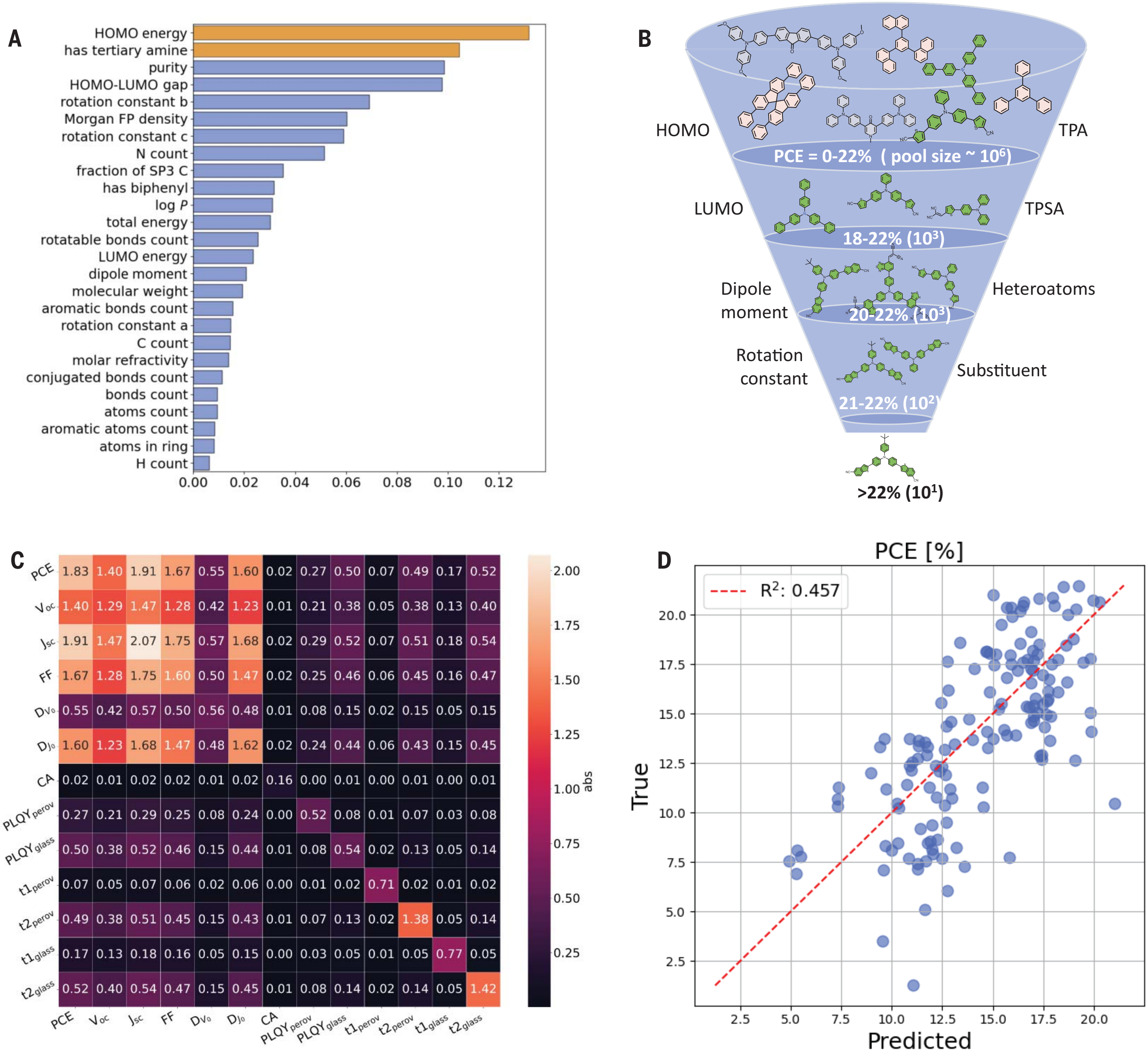


Fig. 5. Model analysis. (**A**) Feature importance of the RFM evaluated from the *M* matrix coefficients. Log*P* is the water octanol partition coefficient, which acts as an indicator of solubility in polar versus nonpolar solvents. Morgan fingerprint (FP) density is an approximative and relative count of the diversity of substructures in a molecule. More symmetric molecules with only few simple building blocks have lower Morgan FP densities compared with complex, nonsymmetric molecules with diverse building blocks. (**B**) Molecular design rules guided by machine learning results. (**C**) Covariance matrix of the multilabel GPR for PCE and additional device characteristics. $t1_{perov}$ and $t2_{perov}$ are the carrier lifetimes of two recombination processes in the glass/HTM/perovskite device, with the laser incident from the perovskite side, whereas $t1_{glass}$ and $t2_{glass}$ are measured with the laser from the glass side. $PLQY_{perov}$ and $PLQY_{glass}$ refer to the photoluminescence quantum yield of the glass/HTM/perovskite device, measured with the laser incident from the perovskite and glass sides, respectively. DV_0 and DJ_0 are the derivatives of the *JV* curve in $V = 0$ and $J = 0$, respectively. CA is the contact angle. Abs, absorbance. (**D**) Linear model test predictions for the leave one out CV scheme.

We adopted a bottom-up optimization approach, from HTM to perovskite to electron-transporting material (ETM) (see section S5 in the supplementary materials). These efforts improved device PCE by 10 to 20%, reaching performance levels of 23.5 to 24.3% with a fill factor (FF) of 87%.

Considering that new conditions might affect the original ranking of materials, thereby influencing ML predictions, we selected 20 molecules from the database to observe the impact of individualized optimization on trends. The results showed that it slightly affected the open-circuit voltage (V_{oc}) and FF trends, but the impact on the PCE trend was negligible (fig. S30). To further fully explore the potential of those materials on V_{oc} and short-circuit current density (J_{sc}), we further refined the perovskite formulation and ETM to better align with the properties of those HTMs (Fig. 4E).

As a result, the reference device using the state-of-the-art small molecule MeO-2PACz achieved a PCE of 24.6%, with a V_{oc} of 1.165 V, J_{sc} of 25.6 mA/cm^2, and FF of 80%. By contrast, when using the same recipe, A1090B769 (sample 130, a small molecule that synthesized in the second iteration) showed a substantial improvement in the performance, achieving a V_{oc} of 1.195 V, a J_{sc} of 26.0 mA/cm^2, and a FF of 84%. These enhancements culminated a PCE of 26.2% (certified 25.9%) alongside an operational stability (ISOS-L2, 65°C with maximum power point tracking, under light) (*33*, *34*) and maintained 80% of the initial performance (T_{80}) for more than 1000 hours. This is attributed to our HTM's passivating properties, suppressing nonradiative recombination at the interface more effectively. On the other hand, the state-of-the-art polymer PTAA demonstrated notably lower performance, with a V_{oc} of 1.105 V, a J_{sc} of 24.0 mA/cm^2, and a FF of 82.4%. Finding performance values as high as 26.2% for a process, significantly surpassing those of state-of-the-art polymers and small molecules, demonstrates the power of this approach. Moreover, it outlines the next major steps for the combined material and device acceleration strategy that replace single-objective optimization with multiobjective optimization tasks. We emphasize that, although this study uses PCE as the sole optimization criterion, stability is regarded as equally important in the field of PSCs. Encouraged by the outcome of this first campaign, we are currently preparing to implement a multiobjective optimization routine for efficiency and stability as a function of the HTM in our workflow. Looking forward in a positive manner, we see material costs, toxicity, and recyclability as potential future optimization criteria to be included in such materials discovery campaigns.

Model analysis

To obtain interpretable insight into what our ML models have learned and to identify physical parameters that influence the device performance, we added further experimental material properties and extracted feature importance information from the trained ML models (Fig. 5). We first focused on analysis of the model to better understand which molecular descriptors are relevant for the model's predictive performance within the generated data. We then systematically evaluated which molecular descriptors are relevant for the model's ability to generalize to unseen building blocks and, thus, new molecules, which is relevant for molecular discovery. Lastly, we focused on the usefulness of additional experimental observations in improving model performance to help find relations between PCE and other device characteristics, which might serve as intermediate measurements in the future to speed up the experimental iteration cycles by introducing proxy measurements or stopping criteria. Because the HOMO position is relevant for charge extraction from physical considerations, the HOMO level position is expected to be aligned between the electrode and the perovskite and is identified as a notable feature in Fig. 5A (*35*–*37*).

To identify more decisive features, we conducted a feature analysis using the Recursive Feature Machine (RFM), a kernel machine that recursively learns features importance (*38*). As in Radhakrishnan *et al*., we used a generalization of the Laplacian kernel that incorporates a learnable feature matrix M to compute the Mahalanobis distance between data points (*39*). The coefficient of determination (R^2) on the test data of this method was evaluated to be approximately 0.5. On 100 randomized train-test splits, we ranked the largest feature matrix values of the trained Kernel methods in Fig. 5C. The features on which the RFM model focused are purity, HOMO level, HOMO/LUMO gap, and the presence of tertiary amines.

Aside from the electronic properties of the molecule, the purity of the synthesis product is the most crucial descriptor for the final device performance. This finding corroborates that impurities typically reduce overall performance because of potential diffusion and the introduction of traps or unwanted doping in adjacent layers. Other important factors were the presence of nitrogen atoms, which reflects the observations mentioned above in Fig. 2F, dipole moment, molecular shape, and, to a lesser degree, composition and overall bond type or conjugation.

Some of those descriptors might hint at underlying physical structure-property-function relations, whereas other descriptors are only relevant for the specific dataset and choice of building blocks. Because the dataset is not strictly independent and identically distributed, we additionally evaluated the generalizability of our ML models in a leave-one-out cross-validation scheme. We iteratively picked each single molecule as a test set and removed the same A and B building blocks from the training set to make sure that the model cannot produce its predictions by simply learning to recognize the molecular fragments. We trained GP and RFM models and evaluated their test error with this validation scheme, reaching in both settings an R^2 value of approximately 0.3. Although the test error increased, the model could still generalize to unseen data points. The most relevant features in this setup remained similar to Fig. 5A in the importance ranking: the presence of tertiary amines, HOMO/LUMO gap, and HOMO level, but also dipole moment and the presence of biphenyl substructures. We confirmed that the model learned to predict the perovskite device performance for unseen HTM molecules, as required for materials discovery. We furthermore confirmed that the structural features generalize to unseen building blocks, which increases the likelihood that they encode underlying physical relationships. However, further expansion of the set of investigated building blocks would be required to further elucidate these relations.

To have a more interpretable model, we trained a linear regression model applying first both a forward and a backward sequential feature selection, which is a family of greedy search algorithms used to reduce the feature space to a lower subset. We evaluated the resulting models with the Bayesian information criterion to select the best-performing set of features we had learned so far. Our selected model uses eight features [aromatic bonds and atoms counts, logarithm of the partition coefficient (log*P*), count of nitrogen atoms, purity, dipole, rotation constant *c*, and the presence of tertiary amines] to predict the PCE, achieving an R^2 of approximately 0.46 (Fig. 5D), which is higher than that of any other model we found.

Additional experimental input from extended characterization was generated to increase the diversity of input and output parameters, including wettability, photoluminescence quantum yield (PLQY), and time-resolved photoluminescence (TRPL) (see sections S1 and S4 of the supplementary materials for details). Because these measurements are often only available after producing the device, we added them as ulterior outputs and trained a MTGPR on the joined target space. The task covariance matrix represents potential correlations that the MTGPR model uses with a kernel function shared among tasks (Fig. 5C). It shows the expected correlation between V_{oc}, J_{sc}, and FF with the PCE. To a lesser extent, it also indicates correlations between PCE and additional labels, for example, with TRPL, which would be consistent with current reports (*40*–*43*). However, the correlation was not as pronounced as expected and did not lead to a statistically significant improvement in PCE prediction accuracy when the additional labels were fed to the multitask model.

Lastly, to provide chemists and material scientists with a clearer understanding of our findings, enabling them to delve deeper into molecular design based on our finds, we used chemical language to elucidate the results of ML (Fig. 5B). The feature importance plot distinctly highlights the significance of HOMO and the presence of tertiary amines in the model. The significance of HOMO in molecular design has been widely recognized, given its decisive role in charge extraction at interfaces. However, the presence of tertiary amines is often overlooked.

Upon examining all synthesized molecules, we discovered that the presence of tertiary amines often refers to TPA, the low ionization potential of which contributes greatly to the molecule's HOMO (*2*). Based on these two descriptors, all synthesized molecules could be categorized into three types: Type I, TPA-absent molecules, referred to as AxBy; Type II, with TPA on the periphery, typically AxB1 structures; and Type III, with TPA at the molecular center, typically A30By structures.

Under this classification, a pattern emerged in HOMO and PCE: (i) HOMO ranges from 5.1

to 6.1 eV with 5 to 14% PCE; (ii) HOMO from 4.3 to 5.2 eV with 13 to 20% PCE; and (iii) HOMO from 4.9 to 5.7 eV with 15 to 21% PCE. This classification narrowed the candidate pool from 9.6×10^5 to 5.8×10^3 molecules. Once the A position was established as TPA, molecular properties were primarily influenced by the B-position group. The feature importance analysis also highlighted the roles of the HOMO/LUMO gap and dipole moment. The combination of TPA and acceptors ensured an appropriate HOMO-LUMO bandgap, with heteroatoms in acceptors also contributing to perovskite passivation. This combination further reduced candidates from 5.8×10^3 to 4.6×10^2 molecules. To facilitate rapid selection by chemists without DFT calculations, we used the topological polar surface area (TPSA) as a rough indicator of the building block's polarity and electron-withdrawing capacity, which is readily searchable with PubChem.

The performance of molecules should be finely tuned based on TPA and acceptors, such as the edge-on orientation positively impacting passivation and charge transport. In the combination of TPA derivative and five B groups (fig. S22), device performance was systematically enhanced through fine-tuning of the B-position group and TPA structure. For example, groups such as A770, with slightly weaker symmetry, tend to exhibit better device performance. This fine-tuning can reduce the number of candidate molecules from 100 to 10, a quantity that is well within the realm of high-throughput synthesis.

We summarize by highlighting the two-fold strategy learned from training an ML model to be capable of predicting such a complex property as a device performance based on molecular structure input. Such models can be further explored in a twofold strategy. On the one hand, it can be used in autonomous workflows to identify and predict further novel molecules. On the other hand, synthetic researchers can use that model to predict perovskite device performance for new molecular designs within a certain chemical space, and that process can be further guided and supported by the set of design rules extracted that are elucidated from a fully trained model.

Conclusions

We demonstrate a workflow for discovery of functional materials optimized for highly complex applications such as photovoltaic devices. We built predictive models based on molecular descriptors, allowing us to link the structure of a material to the performance of a highly complex device, such as a solar cell. The inclusion of organic synthesis in self-driven autonomous labs in combination with autonomous device optimization lines enabled this workflow, and this approach could be extended to other application areas. This capability is particularly important for device processing and optimization, which necessitates a nuanced understanding of both the material and the process involved.

Looking forward, we aim to integrate material discovery and device optimization into a seamless, closed-loop process. Achieving this will require a concerted effort in interdisciplinary research, combining insights from materials science, engineering, and advanced computational techniques to create a synergistic workflow. This integrated approach is the most promising strategy to revolutionize the way we develop and optimize materials for cutting-edge technological applications.

REFERENCES AND NOTES

1. D. Meng *et al*., *Chem. Rev.* **122**, 14954 14986 (2022).
2. J. Wang, K. Liu, L. Ma, X. Zhan, *Chem. Rev.* **116**, 14675 14725 (2016).
3. A. Farokhi, H. Shahroosvand, G. D. Monache, M. Pilkington, M. K. Nazeeruddin, *Chem. Soc. Rev.* **51**, 5974 6064 (2022).
4. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **360**, 186 190 (2018).
5. B. J. Shields *et al*., *Nature* **590**, 89 96 (2021).
6. T. Mou *et al*., *Nat. Catal.* **6**, 122 136 (2023).
7. P. M. Maffettone *et al*., *Nat. Comput. Sci.* **1**, 290 297 (2021).
8. Y. Liu *et al*., *Nat. Mach. Intell.* **4**, 341 350 (2022).
9. A. Merchant *et al*., *Nature* **624**, 80 85 (2023).
10. E. Ozer *et al*., *Nat. Electron.* **3**, 419 425 (2020).
11. H. Yang *et al*., *Nat. Mach. Intell.* **4**, 84 94 (2022).
12. R. Gómez Bombarelli *et al*., *Nat. Mater.* **15**, 1120 1127 (2016).
13. T. C. Wu *et al*., *Adv. Mater.* **35**, e2207070 (2023).
14. S. Langner *et al*., *Adv. Mater.* **32**, e1907801 (2020).
15. X. Du *et al*., *Joule* **5**, 495 506 (2021).
16. J. Zhang *et al*., *Adv. Energy Mater.* **13**, 2302594 (2023).
17. T. Osterrieder *et al*., *Energy Environ. Sci.* **16**, 3984 3993 (2023).
18. J. Xu *et al*., *Nat. Mater.* **22**, 1507 1514 (2023).
19. Y. Bai *et al*., *J. Am. Chem. Soc.* **141**, 9063 9071 (2019).
20. H. Lu *et al*., *Nature* **604**, 662 667 (2022).
21. M. Seifermann, P. Reiser, P. Friederich, P. A. Levkin, *Small Methods* **7**, e2300553 (2023).
22. J. Wu *et al*., *J. Am. Chem. Soc.* **145**, 16517 16525 (2023).
23. P. Yan, D. Yang, H. Wang, S. Yang, Z. Ge, *Energy Environ. Sci.* **15**, 3630 3669 (2022).
24. M. Jeong *et al*., *Nat. Photonics* **16**, 119 125 (2022).
25. X. Yu *et al*., *Angew. Chem. Int. Ed.* **62**, e202218752 (2023).
26. A. F. Zahrt *et al*., *Science* **363**, eaau5631 (2019).
27. W. Yan *et al*., *Adv. Energy Mater.* **6**, 1600474 (2016).
28. J. Urieta Mora, I. García Benito, A. Molina Ontoria, N. Martín, *Chem. Soc. Rev.* **47**, 8541 8571 (2018).
29. S. Grimme, C. Bannwarth, *J. Chem. Phys.* **145**, 054103 (2016).
30. J. Wu *et al*., *Chem. Eng. J.* **422**, 130124 (2021).
31. H. Guo *et al*., *Angew. Chem. Int. Ed.* **60**, 2674 2679 (2021).
32. MatPortal, Thin film solar cell ontology; https://matportal.org/ontologies/TFSCO.
33. J. A. Hauch, C. J. Brabec, N. Fabricius, W. Bergholz, *Energy Technol.* **8**, 2000487 (2020).
34. M. V. Khenkin *et al*., *Nat. Energy* **5**, 35 49 (2020).
35. N. J. Jeon *et al*., *J. Am. Chem. Soc.* **136**, 7837 7840 (2014).
36. N. J. Jeon *et al*., *Nat. Energy* **3**, 682 689 (2018).
37. M. Jeong *et al*., *Science* **369**, 1615 1620 (2020).
38. A. Radhakrishnan, D. Beaglehole, P. Pandit, M. Belkin, Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn featuresarXiv:2212.13881 [cs.LG] (2023).
39. A. Bellet, A. Habrard, M. Sebban, *Springer Nature* **9**, 1 151 (2015).
40. Y. Zhao *et al*., *Nat. Energy* **7**, 144 152 (2022).
41. T. Zhang *et al*., *Science* **377**, 495 501 (2022).
42. X. Li *et al*., *Science* **375**, 434 437 (2022).
43. T. Bu *et al*., *Science* **372**, 1327 1332 (2021).
44. Aimat lab, perovskite_htm_screening, Github (2024); https://github.com/aimat lab/perovskite_htm_screening.

ACKNOWLEDGMENTS

Funding: J.W. and J.L. acknowledge financial support from the Sino German Postdoc Scholarship Program (CSC DAAD). C.J.B. gratefully acknowledges financial support through the "Aufbruch Bayern" initiative of the state of Bavaria (EnCN and "Solar Factory of the Future"), the Bavarian Initiative "Solar Technologies go Hybrid" (SolTech), and the German Research Foundation (DFG) SFB 953 no. 182849149. J.W., J.H., Y.Z., and C.J.B. gratefully acknowledge the grants "ELF PV Design and development of solution processed functional materials for the next generations of PV technologies" (no. 44 6521a/20/4) by the Bavarian State Government. J.Z., Z.X., and K.Z. acknowledge financial support from the China Scholarship Council (CSC). Y.Z. acknowledges the Alexander von Humboldt Foundation for supporting his scientific research during the postdoctoral period (grant no. 1199604). Calculations were performed on the HoreKa supercomputer funded by the Ministry of Science, Research, and the Arts Baden Württemberg. P.R., L.T., and P.F. acknowledge support by the state of Baden Württemberg through bwHPC. L.T. and P.F. acknowledge support by the Federal Ministry of Education and Research (BMBF) under grant nos. 01DM21002A (FLAIM) and 01DM21001B (German Canadian Materials Acceleration Center). J.W. and J.S.R. O gratefully acknowledges the financial support from Solar TAP initiative by the Helmholtz Foundation/BMBF. J.L. gratefully acknowledges financial support from the National Natural Science Foundation of China (grant no. 62104031) and the Technical Field Funds of 173 Project (grant no. 24JJ210663A). M.H., B. w.P., and S.I.S. acknowledge financial support from the Basic Science Research Program (NRF 2018R1A3B105282023) and the Nano & Material Technology Development Program (RS 2024 00449449) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (MSIP). **Author contributions:** J.W., J.H., S.I.S, P.F., and C.J.B. conceived the idea and supervised the project. J.W. performed the high throughput synthesis and fabrication of the perovskite device. L.T., P. R., and P.F. applied the ML algorithms. J.W., M.H., and L.W. fabricated the perovskite device and wrote the manuscript. J.Z., K.Z., and Y.W. characterized the PLQY and TRPL. Z.X. developed algorithms to process UV Vis data. J.S.R. O. provided partial materials. L.W., B. w.P., A.B., Y.Z., J.L., L.L., L. L.D., D.M.G., and M.E.P. O. analyzed the data. All authors reviewed and edited the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Code and data to reproduce the plots shown in the machine learning section of the manuscript are publicly available at GitHub (*44*).