Data-based *Polymer-Unit Fingerprint (PUFp)*: A Newly Accessible Expression of Polymer Organic Semiconductors for Machine Learning

Xinyue Zhang¹, Genwang Wei¹,², Ye Sheng³, Jiong Yang³,⁴, Caichao Ye¹,²,*, Wenqing Zhang¹,²,*

¹ Department of Physics & Guangdong Provincial Key Laboratory of Computational Science and Material Design, Southern University of Science and Technology, Shenzhen, 518055, PR China

² Academy for Advanced Interdisciplinary Studies & Department of Materials Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, PR China

³ Materials Genome Institute, Shanghai University, Shanghai, 200444, PR China

⁴ Zhejiang Laboratory, Hangzhou, 311100, PR China

*Corresponding Authors: C. Ye: yecc@sustech.edu.cn, W. Zhang: zhangwq@sustech.edu.cn.

Abstract

In the process of finding high-performance organic semiconductors (OSCs), it is of paramount importance in material development to identify important functional units that play key roles in material performance and subsequently establish substructure-property relationships. Herein, we describe a polymer-unit fingerprint (PUFp) generation framework. Machine learning (ML) models can be used to determine structure-mobility relationships by using PUFp information as structural input with 678 pieces of collected OSC data. A polymer-unit library consisting of 445 units is constructed, and the key polymer units for the mobility of OSCs are identified. By investigating the combinations of polymer units with mobility performance, a scheme for designing polymer OSC materials by combining ML approaches and PUFp information is proposed to not only passively predict OSC mobility but also actively provide structural guidance for new high-mobility OSC material design. The proposed scheme demonstrates the ability to screen new materials through pre-evaluation and classification ML steps and is an alternative methodology for applying ML in new high-mobility OSC discovery.

**Keywords:** Polymer unit, Fingerprint, Machine learning, Organic semiconductors, High mobility
Table of Contents
1. Introduction

Organic semiconductors (OSCs) are widely applied in a variety of areas, such as photovoltaic materials, field-effect transistors and thermoelectric materials, due to their characteristics of low synthesis costs, light weight structures, and strong flexibility and plasticity attributes. Because many properties of OSCs are related to carrier transport, the design and synthesis of novel materials with high carrier mobility has become a key factor in the development of high-performance optical and electrical devices. Semiconductors can be divided into n-type, p-type and bipolar materials; the carriers of p-type semiconductors are holes, and those of n-type semiconductors are electrons. Recent studies have shown that the mobility of polymer semiconductors can be improved by adjusting branched chains, combining different organic functional groups (D-A, D-D, A-A, etc.) or doping with small molecules. The carrier mobility of p-type semiconductor materials has been reported to be as high as 71 cm²·V⁻¹·s⁻¹, and the mobility of n-type semiconductors can also reach 14.4 cm²·V⁻¹·s⁻¹. In recent years, the mobility records of OSCs have continued to advance, which proves that there is plenty of room for improvement in the mobility of OSCs.

Machine learning (ML) can train advanced algorithms based on big data to capture intrinsic connections that may embody underlying principles; thus, ML algorithms have led to significant progress in many fields. In materials research and development, including studies on catalyses, batteries and OSCs, ML is able to construct structure-property material atlases and to design new materials, especially in fields where properties and structures are closely related. When ML is used for the construction of structure-property material atlases, the way in which structural information of materials is expressed determines which material descriptors or features can be studied. Recently, scientists have commonly used molecular fingerprints, the most convenient digital descriptor converted from a string of certain chemical structures (simplified molecular input line entry system: SMILES) using a tool equipped with a computer language platform, to express the structure of organic molecules for ML. In 2019, Sun et al. explored several ways to express molecular structures, i.e., images, ASCII strings, descriptors, and fingerprints, as inputs for various ML algorithms and studied their effect on the prediction of power conversion efficiency (PCE), which is a key performance indicator of optoelectronic materials. It was found that a hybridization fingerprint with a length of over 1000 bits combined with a classification model trained by the random forest (RF) algorithm can obtain high prediction accuracy (up to 81.76% N/P classify - N-type: electronic mobility > hold mobility; P-type: electronic mobility < hold mobility). This high accuracy indicates that molecular fingerprints, as inputs for the expression of molecular structures used in ML, have good predictive performance for organic polymer materials.

Molecular fingerprints, such as FP2, Morgan and Torsion, have been designed for large-scale database screening and take the form of an array of bits. They contain “1”s and “0”s to describe the presence or absence of particular substructures/patterns in a molecule and can be used as inputs to train ML models to predict the properties of molecular materials. Morgan fingerprints are circular fingerprints that are generated by the Morgan algorithm and represent substructures within a certain distance from the target atom. (Figure 1a). The Torsion fingerprint (Figure 1b) is a path-based fingerprint, where each of the 32 digits represent a
linear path of 4 nonhydrogen atoms, and contains π electron numbers, atom types, and nonhydrogen branch chain numbers. The FP2 molecular fingerprint is also a path-based molecular fingerprint that indexes small molecular fragments based on a linear path of seven atoms. However, the direct interpretation of these ML models that use these fingerprints as input is difficult, and the specific substructures that greatly influence the performance cannot be well identified. If each bit of a molecular fingerprint represents a specific substructure, important substructures that greatly influence the target performance can be identified by analysis and reverse-trained ML models; for example, MACCS and La FREMD fingerprints implemented by RDKit can be used to identify key substructures, applied in organic photovoltaics (OPV) material design, and screened by ML. However, the disadvantage of the MACCS fingerprint is that each bit represents a very small chemical structure (such as a double bond, a hydroxyl group, a carbonyl group, an S–N group, or an alkaline metal), and for the La FREMD fingerprint, each bit represents a randomly cropped molecular fragment according to certain rules. Obviously, if substructures are not a common functional structure, such as polymer units in OSCs, it is difficult to apply the identified key substructures to construct new materials and explore the relationship between structure and properties.

**Figure 1.** Schematic diagrams of several molecular fingerprints. (a) **Morgan:** Morgan fingerprints consist of hash indexes of substructures within a certain distance of each atom in the structure. (b) **Torsion:** Torsion fingerprints are linear paths of 4 atoms at a time, and the number of π electrons, atomic species, and nonhydrogen branch chains of these 4 atoms are recorded in the fingerprints. (c) **MACCS:** Each bit in the MACCS code represents a small substructure and adopts one-hot encoding. (d) **La FREMD:** The La FREMD fingerprint generated by RDKit is a molecular fragment generated by clipping the input according to certain rules. One-hot coding was used to represent each molecular fragment.

In this work, a data-based dynamic polymer-unit fingerprint (PUFp) was designed and proposed to
segment suitable functional structure units to represent each bit of a molecular fingerprint. All organic monomers are bonded by polymer units (mono rings, bicyclic rings, fused rings and branched chains). Therefore, although organic monomers are multitudinous, polymer units are common and restricted. The PUFp generation framework is schematically illustrated in Figure 2. The organic polymers and units are dynamically generated according to a user-provided database through the Python-based polymer-unit-recognition script (PURS, access by this link: https://github.com/xinyue123-q/Python-based-polymer-unit-recognition-script-PURS-) or see the Supporting Information), and each bit represents a polymer unit. We demonstrate that PUFp can be utilized in the entire process during the investigation of polymer OSCs by ML. The N/P classification accuracy of an OSC is above 80%. Through the Shapley Additive exPlanations (SHAP) analysis of trained ML models, the key polymer units that greatly influence the mobility of OSCs could be identified. A polymer-unit library that consists of 445 units is created, and the top-positive-important units for n- or P-type mobility and N/P classification are discovered. The relationship and mechanism between the OSC polymer units and their combination modes with experimental carrier mobilities are also investigated and studied. Through this work, the application mode of fingerprints changes from “black” to “white”, i.e., switching from a passive “black box” model construction to active mechanism exploration and a new material design. As a result, the ML method combined with PUFp can fundamentally change the methodology for new organic material discovery.

2. Methods

2.1 Summary of the polymer-unit fingerprint (PUFp) generation framework for developing high-mobility OSC materials with ML.

The PUFp generation framework and the workflow for its application in developing high-mobility OSC materials with ML are schematically illustrated in Figure 2. It consists of six integral components that perform four tasks separately:

I. Starting with the establishment of a polymer OSC materials dataset, which consists of 678 different polymer OSC materials that have been experimentally reported (Figure 2a and SI 1. Polymer-unit SMILES). The N/P type distribution of these data is shown in Figure S1. We analysed the SMILES of each OSC and divided each polymer unit by using our PURS framework (Figure 2b), and the polymer-unit library was created (Figure 2c). Here, the polymer units are updated according to the polymer OSC database, so the polymer unit library is updated dynamically based on the OSC database.

II. Each polymer OSC SMILES in the polymer OSC dataset (Figure 2a) was indexed by the polymer-unit library (Figure 2c) to generate PUFp in one-hot form (Figure 2d). First, a node matrix is generated according to the polymer-unit index in the database and the polymer-unit library, and each node can contain one polymer unit. The number of columns in the node matrix represents the maximum number of nodes $N$ according to the OSC database, and the number of rows represents the number of polymer-unit types $T$.
according to the polymer-unit library. Then, we sum each row of the node matrix and obtain a one-dimensional vector, which we named PUFp.

III. Five commonly used ML algorithms are employed by the self-built polymer OSC materials database and our newly designed PUFp framework, which can express 445 different polymer units of an organic molecule, to implement property prediction model construction and unit importance analysis. With this targeted feature engineering, we can identify the key basic polymer units that greatly influence the OSC mobility performance (Figure 2e).

IV. A new material virtual library is generated by rearranging the most important polymer units based on the three main types of polymer units (mono rings, bicyclic rings, fused rings and branched chains; see Figure 2f). The ML model trained in the previous step is used to predict the mobility performance of the new material virtual library, and the new materials are classified into high, medium, and low types. The polymer-unit type combinations with the highest proportion of the high-mobility types are calculated, and the favoured polymer-unit combinations for high-mobility OSCs are found.

Figure 2. Scheme of the PUFp generation framework and the workflow for its application in developing high-mobility OSC materials with ML.

2.2 Polymer-unit library

The polymer unit identified by PURS is collected and added to the polymer-unit library. The uniqueness of the polymer unit is ensured, and it is classified by a Python script (Figure 3a, SI: 1. Polymer-unit SMILES and 2. Polymer-unit types). First, the units can be divided into two categories: branch chain and ring units (Figure 3b). Branch chains are classified by the elements they contain (Figure 3c). Ring units are first classified by the number of rings and then by the element. Through this classification approach, the polymer-unit grouping rules of materials with high mobility can be analysed. Moreover, new materials can be obtained by combining different types of polymer units.
2.3. Procedure for acquiring the polymer OSC database & generation strategy for a polymer-unit fingerprint (PUFp)

The polymer OSC database contains 678 SMILES codes (standardized by RDKIT) of the polymer OSC materials obtained from published references, and each data point was labelled with \( N \) or \( P \) carrier mobility labels (\( \text{cm}^2\cdot\text{V}^{-1}\cdot\text{s}^{-1} \) as the unit) (Figure 2a). Here for clarity, it should be stated that whether a polymer is \( N \) or \( P \) type depends on the electrode identity as well as the chemistry, and the collected experimental data is hard to standardized to all have the same electrode structures. Although the data from different literatures have different experimental conditions, there must be some meaningful correlation between the size of the mobility and the structure of the polymer. The distribution of \( N \)-type and \( P \)-type data and its influence are discussed in Fig.S2-S3. Polymer units are considered the basic functional building blocks for polymer molecular structure construction. An in-house designed script called PURS was prepared to recognize polymer units and generate corresponding fingerprints according to the standardized SMILES codes. The PURES program can identify and divide all the polymer units in the SMILES of each OSC. PURES identifies and divides polymer units from the input SMILES codes following these rules:

1. By dividing molecules at their breakpoints, polymer OSC molecules are separated into several functional polymer units.

2. The breakpoint must be a single bond that connects two independent units (mono ring, bicyclic ring, fused ring and branched chain units\(^1\)).

---

\(^1\) The carbon chain skeleton of a polymer monomer is the first one connected, called the main chain, and the carbon chain outside the main chain is the branch chain.
3. The divided linear segments that are acyclic structures are uniformly referred to as chains, and if the segment contains less than 5 atoms in a chain, these chain segments should be combined with the neighbouring ring unit to form a new larger unit. Since there are few acyclic segments with lengths that are greater than 5 atoms on the main chain of an organic semiconductor, the other long-chain unit can be considered a branched chain.

4. All polymer-unit SMILESs need to be canonicalized by RDKIT to ensure that the same polymer unit corresponds to the same SMILES code.

After all polymer units are collected and repeated data are removed, a complete “polymer-unit library” is summarized (Figure 2c). Here, if the number of polymer units in the polymer-unit library is T and the maximum number of polymer units in the data are N, then we assume a “node matrix” in which the numbers of rows and columns are T and N, respectively, exists, and “1”s and “0”s are used to describe the presence or absence of particular polymer units in an OSC molecule. Obviously, the node matrix expresses that each OSC data point is composed of polymer units, and the matrix expresses the number of each polymer unit. Finally, each row of the node matrix is summed to generate a one-dimensional vector/fingerprint, named PUFp. In this work, the 678 polymer OSC SMILESs in the polymer OSC database (Figure 2a) were indexed by a polymer-unit library to generate the corresponding PUFp model in one-hot form (Figure 2d) 58, as PUFp contains information about the type of polymer units and the number of each polymer unit. The length of the generated fingerprint is 445 bits, which means that the 687 OSC data points in this work are composed of 445 different polymer units.

2.4 ML algorithms & statistical analysis based on PUFp

With PUFp as input, five ML algorithms, the RF 59, K-nearest neighbour (KNN) 60-61, Gaussian process 62, support vector machine (SVM) 63, and multilayer perceptron (MLP) models 64, are implemented with Python scripts using scikit-learn 65 to test the classification performance of different algorithms (Figure 2e). First, the N/P classification of polymer OSC data is performed by the ML models. n- or p-type OSCs are defined as follows: if the electron mobility is greater than the hole mobility, the OSC is considered to be an n-type OSC; otherwise, it is a p-type OSC. Then, according to the experimental mobility values in the polymer OSC database, the mobility classification models for n- or p-type polymer OSCs are trained and tested by the five ML algorithms mentioned above. The mobility is classified as high (> 4 cm²·V⁻¹·s⁻¹), medium (1 to 4 cm²·V⁻¹·s⁻¹), or low (0 to 1 cm²·V⁻¹·s⁻¹). According to the above classification, there are 477 p-type semiconductors, 207 n-type semiconductors, 107 high mobility data, 212 medium mobility data, and 500 low mobility data in the polymer OSC database.

PUFp can not only be used as input for ML but also be used to perform statistical analysis on the database (Figure 3). The types and number of polymer units in the database are analysed, and the polymer-unit combinations of OSCs with high mobility are explored. The best ML model among the five algorithms is used to screen the new materials (See Fig.S1 for more machine learning results on PURS). We further predict the
potential mobility values of the newly designed polymer OSC materials based on the combination of the selected key basic polymer units that greatly and positively influence the OSC’s mobility performance. Compared to the published materials with the same type of polymer units, the rationality of these high-mobility polymer-unit combinations is verified.

3. Results and discussion

3.1 Reliability of PUFp as a structural expression of the polymer OSC molecule in ML modelling.

![Figure 4](image)

**Figure 4.** Performance of the ML models (MLP, KNN, Gaussian, SVM and RF) using PUFp as input.

The PUFp model (445 bits in this work) is used as input to train and test the classification performance of the MLP, KNN, Gaussian, SVM and RF models. The GridSearchCV is used to adjust the parameters of each algorithm, and 10-fold cross-validation are performed, see Table S1-Table S3. As shown in Figure 4, the N/P classification accuracy of all five algorithms can reach more than 80%. Then, based on the database, the mobility of n- or p-type OSCs is classified into three categories: high (> 4 cm²·V⁻¹·s⁻¹), medium (1 to 4 cm²·V⁻¹·s⁻¹) or low (0 to 1 cm²·V⁻¹·s⁻¹). The mobility classification test accuracy of both algorithms for n- and p-type OSCs is more than 70%. We note that among the five algorithms, MLP algorithm achieved the highest prediction accuracy of 75.1% for n-type OSCs, while SVM and RF algorithm were both reach 79.1% prediction accuracy for p-type OSCs. The highest accuracy of N/P classification was 85.2%, which came from MLP, slightly higher than RF algorithm (84.0%).
In addition to FUFp, MACCS(166bits) and Morgan(1023bits) molecular fingerprints are molecular fingerprints with structure for each key. Figure 5 (a-c) is a comparison of these three fingerprints. In general, there was little difference in accuracy between the three fingerprints. In P-type classification, FUFp performs more prominently in SVM and RF algorithms; In N type classification and N/P classification, the best performance is Morgan fingerprint. The ambipolar of OSC is also a topic of considerable concern, so ambipolar is added to the N/P classification (Figure 5-d, the definition of bipolar OSC in this work is: N-type and P-type carrier mobility greater than 1 cm²v⁻¹s⁻¹). It is found that the accuracy of N/B/P classification is slightly lower than that of N/P classification, but it still has 82.7% accuracy under MLP algorithm.

From the results described above, we can conclude that, generally, these algorithms are based on different rules of statistical analysis but can establish accurate relationships between the chemical structure and mobility when PUFp is used as input for building ML models to predict mobility because PUFp is a notable structural expression of polymer molecules that is easily accessible and contains important chemical information (the type of polymer units and the number of each polymer unit of the polymer molecule).

### 3.2 Identification of key polymer units that greatly influence OSC mobility by ML.
Figure 6. Identification of key polymer units by using a SHAP evaluation. (a) The generation strategy of PUFp; (b, c, d) The interpretations of the ML models for n-type, p-type and N/P classification by the SHAP evaluation. The blue and red bars on the right denote the proportional relation between the units and the prediction values. (e, f, g) Chemical structures of the polymer units and their roles are identified through the importance analysis.

The RF method performs the well both in N/P-type classification and mobility prediction of polymer OSCs (Figure 4) because its strategy is to choose multiple features rather than all features from the input for establishing a relationship, which is advantageous when dealing with complex and long input. Therefore, we performed a SHAP evaluation based on the RF model to identify the key polymer units. SHAP is a feature importance analysis method based on a tree model (the RF model is used for this work). The SHAP value can be positive or negative and reflects the contribution of the corresponding feature to the specified category of data. The SHAP evaluation was carried out for n- or p-type high-, medium- and low-mobility classification and N/P classification, respectively, with PUFp as input. The polymer-unit features (Figure 6 b, c, d) with relatively large SHAP values in the three RF models (n-type, p-type, N/P classification) were analysed and labelled as important or favoured polymer units.

OSCs constructed by favoured polymer units are more likely to be classified into a specified category (for example, OSC structures with favoured polymer units positive for n-type mobility are more likely to be classified as n-type high-mobility OSCs). For n-type OSCs, the favoured polymer units are listed in Figure...
polymer units contain imide groups, the electron-withdrawing groups that contribute to lowering the LUMO level of the polymer and facilitating electron injection into the conduction band \(^67\). The main function group of the No. 431, No. 243, and No. 361 polymer units is isoindigo \(^68\). The structure of isoindigo has good plane properties and a strong electron absorption ability, isoindigo is always used as an electronic Acceptor in D-A or A-A polymer OSCs, and isoindigo favours a reduction in the \(\pi-\pi\) stack distance.

For \(p\)-type OSCs, the favoured polymer units are listed in Figure 6f. Polymer unit No. 110 is a full carbon fused ring with large conjugate \(\pi\) bonds, and the complanate delocalized orbitals are beneficial to carrier transport. Polymer unit No. 224 contains a thiazole structure \(^69\), and the electrostatic attraction between the sulphur and nitrogen atoms in thiaoles is beneficial to forming a closer \(\pi-\pi\) packing structure, which is a common strategy in the D-A OSC design \(^70\)–\(^71\).

Figure 6g shows the favoured polymer units for \(n\)-type OSCs in the N/P classification task. Here, No. 284 and No. 7 are both polymer units with thiazole structures. This result is in agreement with the results in previous work, in which the low reorganization energy of thiazole can promote the hopping of electrons between units in OSC molecules \(^53\), and the electron affinity of thiazole units is always higher than 3 eV and can reduce the barrier for electron injection \(^72\)–\(^75\). Polymer unit No. 423 is a BTT structure and is an effective accepter polymer unit in \(n\)-type OSCs. Unit No. 38 with nitrogen, unit No. 85 with cyanide and unit No. 106 with aldehyde are all modified electron-withdrawing groups, and their electron-deficient effect is beneficial for improving the air stability of \(n\)-type semiconductors \(^76\)–\(^78\). In addition, the remaining polymer units in Figure 6e to g (No. 431, No. 243, No. 361, No. 178, No. 137, No. 280, No. 110, and No. 224) are commonly used \(\pi\)-spacer or side-chain structures in OSC molecules.

3.3 Exploration of favoured polymer-unit combinations for the mobility of OSCs

To explore the relevance of polymer-unit combinations to OSC mobility, we first classify the polymer units into three groups (mono ring, bicyclic ring or fused ring, and branched chain groups) according to the backbones that frequently appeared in high-mobility OSCs, i.e., D-A, D-D, and A-A with branched chains. The classification and abbreviation rules for the polymer units are defined in Figure 7c and are named according to the polymer-unit groups (mono ring, bicyclic ring or fused ring, and branched chain groups) and the element modification approach (the formula of modification elements). Then, the combinations of the mono ring, bicyclic ring or fused ring, and branched chain groups in each OSC of the database were counted by statistically analysing the PUFp models. There were 19 kinds of branched-chain combinations, 28 kinds of mono-ring combinations and 75 kinds of bicyclic- or fused-ring combinations. For the details of each combination category, please see SI (SI 2.3 Fused/bicyclic ring combination). A ternary map algorithm was performed and the mono ring, bicyclic ring or fused ring, and branched chain categories were used as the axes for all OSCs in the studied database; see Figure 7a and Figure 7b. It is found that the OSCs in the studied database are concentrated around several certain combinations. The ternary map has many empty areas in both the \(n\)- and \(p\)-type OSCs, and obviously, these unreported combinations generate a huge number of structures.
(19×28×75=39900). To reduce the number of unreported candidates OSC materials, the scope of the mono ring, bicyclic ring or fused ring, and branched chain group is limited to polymers with high mobility. A scheme for high-mobility combination prediction with ML is shown in Figure 7d. The OSC monomer consists of a mono ring, bicyclic ring or fused ring, and branched chain. The Cartesian product of the mono ring-type combination, bicyclic ring or fused ring-type combination and branched chain-type combination in the search space is taken to form a polymer-unit type combination of OSCs. Each combination generates 500 pieces of data. The carrier mobility of these data is predicted using the trained RF model shown in Figure 4. The ratio of predicted high-mobility data among 500 data points for each combination was calculated, and the combinations with the highest ratios were found.

![Diagram](image)

**Figure 7.** Exploring favoured polymer-unit combinations for the mobility of OSCs. (a, b): Ternary distribution of mobility with different combinations of polymer units for n- and p-type OSCs in the studied database, with the categories of mono ring, bicyclic ring or fused ring, and branched chain as axes. (c) Classification and abbreviation definitions for the polymer units. (d) Scheme of high-mobility combination prediction by ML.

According to the high-mobility combination predicted with ML shown in Figure 8d, the favoured polymer-unit combinations for high-mobility OSCs and the top five n-type and p-type polymer-unit
combinations with high-mobility OSCs predicted with ML are listed in **Figure 8l** and **Figure 8m**, respectively. **Figures 8a to 8k** show the high-mobility polymer OSC structures ($\mu > 4 \text{ cm}^2\cdot\text{V}^{-1}\cdot\text{s}^{-1}$) reported in published works, and in these polymer structures, one or more polymer units are of the same type as those in the polymer-unit combination shown in **Figure 8l** and **Figure 8m**. The boundary colours of the polymer-unit combination types in **Figure 8l** and **Figure 8m** are set the same as the colour of the corresponding polymer-unit structure in **Figure 8a to 8k**.

**Figure 8.** Favoured polymer-unit combinations with high-mobility OSCs. (a-k): Sample high-mobility OSC molecules of the same type of polymer-unit combinations $^{18, 47, 50, 53}$. (l, m): Top five $n$-type and $p$-type polymer-unit combinations with high mobility OSCs predicted with ML.

(这部分内容移至结论部分)As shown in **Figure 7l** and **Figure 7m**, among the top five $n$-type and $p$-type polymer-unit combinations of high-mobility OSCs predicted by ML, eight combinations are multiple-D-A
combinations, indicating that the synergy between multiple $D$-$A$ polymer units is advantageous for mobility 79-82. Due to the large dipole moment between $D$-$A$ polymer units, such as isopindus 11 or DPP structures 14 as acceptor units and thiophene as donor units, the $\pi$-$\pi$ stacking 50, 53 distance between polymer molecular layers is small (Figure 7d and g), and the mobility is improved. Moreover, the dihedral angles between adjacent polymer units also affect the carrier mobility 13, and improving the coplanar properties between polymer units is the key to increasing the mobility. Selecting suitable polymer units can improve the coplanar properties. For example, the combination of thiazole (acceptor) and CDT (cyclopentadithiophene, donor, Figure 7a, b) or IDSe (indacenodiselenophene, Figure 7c) has good coplanar properties 11. Figure 7-l,m summarize the top five Polymer-Units combinations with high mobility ratio of type N and type P. In these combinations, all have fluorinated side-chain. In addition, 9 of the 10 combinations have fluorinated substitution structures such as single ring containing fluorine. It fully illustrates the importance of fluorine substitution in improving mobility.

Close intermolecular packing is an important factor for the efficient charge transport of OSCs. Branched chain engineering, such as adjusting the length, shape, and modification approach, can not only increase the solubility of an OSC polymer but can also effectively reduce the $\pi$-$\pi$ stacking distance. Semi or fully fluorinated branch chains can strengthen the rigidity and improve the self-assembly effect of OSC materials 50. As shown in Figure 7l, m, all top polymer-unit combinations of high-mobility OSCs have fluorinated branched chains. This is because fluorinated branched chains (ch-F) can improve the regularity of crystallization. In addition, most of the combinations contain fluorinated units (M-S-F, F-n-O-F) because using fluorine to modify a polymer unit can increase the intermolecular force and improve the coplanar properties of the OSCs (Figure 7d, f, g).

Modulating the electron-donating or electron-withdrawing strength of a repeating unit in $D$-$A$ combinations results in an optimum frontier molecular energy level that can improve the carrier mobility 11. The molecular energy levels of $D$-$A$ combination OSCs can be tuned by substituting appropriate groups in the polymer units with strong electron-withdrawing elements (fluorine, pyridine, nitrile, etc.). This approach lowers the LUMO energy level of the conjugated polymer to improve electron transport. In addition, substituting appropriate groups in the polymer units with strong electron-withdrawing elements (such as fluorine) can change the shapes of HOMO and LUMO along the conjugated backbone of OSC, and the appropriate unit modifications and combinations can improve the intermolecular packing and wavefunction overlap, which also benefit the mobility (Figure 7 d, e, f, g)13. The 10 combinations of Figure 7 (l, m) are analyzed. Each combination consists of three components: branched chain, mono-ring, bicyclic- or fused-ring. Among them, branched chains are used to regulate polymer stacking structure, mono-rings are mostly used as $\pi$ bridges, and bicyclic- or fused-rings are electron donors or acceptors. By comparing N-type and P-type combinations (Figure 7 l, m), it is found that the single rings in P-type combinations are fluorinated sulfur-containing single rings (i.e. fluorothiophene). Among the N-type combinations, the polymer units consists of both sulfur mono-ring (thiophene) and fluoro-sulfur mono-ring (fluorinated thiophene). In other words, the
proportion of fluorothiophene can be higher in P-type combinations, because P-type polymer semiconductors usually require a low HOMO level (-5 eV)\textsuperscript{83}, and fluorine plays a role in lowering the HOMO level.

4. Conclusion:

In summary, based on a database containing realistic polymer OSC materials collected from the literature, an in-house designed script called PURS was demonstrated to recognize polymer units and generate corresponding fingerprints with the PUFp approach, which is a newly developed programming language for expressing OSC molecules and provides distinct results and easy access. A polymer-unit library that consists of 445 units is constructed. Through SHAP, the key polymer units for the mobility of OSCs have been identified, and the top-positive-important units for N- or P-type mobility and N/P classification are discovered.

The purpose of this study is to promote PURS program, and use the PUFp generated by PURS combined with data analysis method to analyze the mobility of organic polymer semiconductors.

The PURS program combined with data analysis methods can achieve: (Reviewer 2 Q13)

1. By using PURS, each polymer-unit in a Polymer monomer can be identified and a polymer-Units database can be created.

2. PURS can be generated by PURS program. Using PUFp as the input feature, the trained machine learning model can achieve more than 80% accuracy in the polymer semiconductor database in this paper. PURS are therefore practical molecular fingerprints. And because each key of the PUFp represents a specific polymer-unit, it is highly interpretable.

3. The machine learning results of PUFp were analyzed using SHAP to identify the polymer-unit that is conducive to the increase of P-type and N-type mobility, and the polymer-unit that is conducive to the formation of N-type polymer semiconductors.

4. 10 combinations of high mobility are summarized. The high mobility polymer search space is obtained by analyzing the existing high mobility polymer combination types. PUFp were generated for the structures in this space, and the ratio of structures predicted to be highly mobile among these structures was counted (using the model trained in 2).

We developed a scheme to help polymer OSC material design by combining ML approaches and PUFp as input not only to passively predict OSC mobility but also to actively provide structural guidance for new high-mobility OSC material design. That is, a large number of polymer OSC materials could be screened through a pre-evaluation and classification process by using our methods, and then the identified leading candidates could be synthesized and further tested by experiments.

Overall, our framework for PURS generation and its application with ML is verified to possess the ability to design high-mobility polymer OSC materials. It is believed that when coupled with the Materials Genome
Initiative and ML, this framework will bring even greater changes to high-mobility polymer OSC material discovery.

Acknowledgements

Funding: Financial support was provided by the Natural Science Foundation of China (92163212), Fundamental Research Program of Shenzhen (JCYJ20190809174203802), Guangdong Provincial Key Laboratory of Computational Science and Material Design (2019B030301001), Guangdong Provincial Young Innovative Talents (2019KQNCX136) and the Key Research Project of Zhejiang Lab (2021PE0AC02). W. Zhang acknowledges support from the Guangdong Innovation Research Team Project (2017ZT07C062). Computing resources were supported by the Center for Computational Science and Engineering at Southern University of Science and Technology.

Author contributions: C. Ye and W. Zhang formulated this project. X. Zhang performed data collection, program coding, and ML analysis. G. Wei and Y. Sheng provided helpful discussion for program coding. X. Zhang and C. Ye cowrote the manuscript. W. Zhang revised the manuscript. W. Zhang and C. Ye secured the funding.

Competing interests: None declared.

Data and materials availability: All data and model analyses are available in the supplementary materials.

References:

1. Yu, G.; Gao, J.; Hummelen, J. C.; Wudl, F.; Heeger, A. J., POLYMER PHOTOVOLTAIC CELLS - ENHANCED EFFICIENCIES VIA A NETWORK OF INTERNAL DONOR-ACCEPTOR HETEROJUNCTIONS. Science 1995, 270 (5243), 1789-1791.

2. Li, G.; Shrotriya, V.; Huang, J. S.; Yao, Y.; Moriarty, T.; Emery, K.; Yang, Y., High-efficiency solution processable polymer photovoltaic cells by self-organization of polymer blends. Nature Materials 2005, 4 (11), 864-868.

3. Shi, L.; Guo, Y.; Hu, W.; Liu, Y., Design and effective synthesis methods for high-performance polymer semiconductors in organic field-effect transistors. Materials Chemistry Frontiers 2017, 1 (12), 2423-2456.

4. Nketia-Yawson, B.; Noh, Y.-Y., Recent Progress on High-Capacitance Polymer Gate Dielectrics for Flexible Low-Voltage Transistors. Advanced Functional Materials 2018, 28 (42).

5. Lee, W. H.; Park, Y. D., Organic Semiconductor/Insulator Polymer Blends for High-Performance Organic Transistors. Polymers 2014, 6 (4), 1057-1073.

6. Lee, J., Physical modeling of charge transport in conjugated polymer field-effect transistors. Journal of Physics D- Applied Physics 2021, 54 (14).
7. Nielsen, C. B.; Turbiez, M.; McCulloch, I., Recent advances in the development of semiconducting DPP-containing polymers for transistor applications. *Adv Mater* **2013**, *25* (13), 1859-80.

8. Li, Y.; Sonar, P.; Murphy, L.; Hong, W., High mobility diketopyrrolopyrrole (DPP)-based organic semiconductor materials for organic thin film transistors and photovoltaics. *Energy & Environmental Science* **2013**, *6* (6), 1684-1710.

9. Xiao, Y.; Zeng, D.; Mazur, L. M.; Castiglione, A.; Lacaze, E.; Heinrich, B.; Donnio, B.; Kreher, D.; Attias, A.-J.; Ribierre, J.-C.; Mathevet, F., A new class of nanostructured supramolecular organic semiconductors based on intertwined multi-lamellar co-assemblies in π-conjugated liquid-crystalline side-chain polymers. *Polymer Journal* **2016**, *49* (1), 31-39.

10. Yang, Y.; Liu, Z.; Zhang, G.; Zhang, X.; Zhang, D., The Effects of Side Chains on the Charge Mobilities and Functionalities of Semiconducting Conjugated Polymers beyond Solubilities. *Adv Mater* **2019**, *31* (46), e1903104.

11. Kim, M.; Ryu, S. U.; Park, S. A.; Choi, K.; Kim, T.; Chung, D.; Park, T., Donor–Acceptor-Conjugated Polymer for High-Performance Organic Field-Effect Transistors: A Progress Report. *Advanced Functional Materials* **2019**, *30* (20).

12. Yoo, S. J.; Kim, J. J., Charge transport in electrically doped amorphous organic semiconductors. *Macromol Rapid Commun* **2015**, *36* (11), 984-1000.

13. Sui, Y.; Deng, Y.; Du, T.; Shi, Y.; Geng, Y., Design strategies of n-type conjugated polymers for organic thin-film transistors. *Materials Chemistry Frontiers* **2019**, *3* (10), 1932-1951.

14. Zou, X.; Cui, S.; Li, J.; Wei, X.; Zheng, M., Diketopyrrolopyrrole Based Organic Semiconductor Materials for Field-Effect Transistors. *Frontiers in Chemistry* **2021**, *9*.

15. Quinn, J. T. E.; Zhu, J.; Li, X.; Wang, J.; Li, Y., Recent progress in the development of n-type organic semiconductors for organic field effect transistors. *Journal of Materials Chemistry C* **2017**, *5* (34), 8654-8681.

16. Paterson, A. F.; Singh, S.; Fallon, K. J.; Hodsden, T.; Han, Y.; Schroeder, B. C.; Bronstein, H.; Heeney, M.; McCulloch, I.; Anthopoulos, T. D., Recent Progress in High-Mobility Organic Transistors: A Reality Check. *Advanced Materials* **2018**, *30* (36).

17. Nielsen, C. B.; McCulloch, I., Recent advances in transistor performance of polythiophenes. *Progress in Polymer Science* **2013**, *38* (12), 2053-2069.

18. Kim, M.; Ryu, S. U.; Park, S. A.; Choi, K.; Kim, T.; Chung, D.; Park, T., Donor-Acceptor-Conjugated Polymer for High-Performance Organic Field-Effect Transistors: A Progress Report. *Advanced Functional Materials* **2020**, *30* (20).

19. Gao, X.; Zhao, Z., High mobility organic semiconductors for field-effect transistors. *Science China-Chemistry* **2015**, *58* (6), 947-968.

20. Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533* (7601), 73-+.

21. Mao, Y.; Yang, H.; Sheng, Y.; Wang, J.; Ouyang, R.; Ye, C.; Yang, J.; Zhang, W., Prediction and Classification of Formation Energies of Binary Compounds by Machine Learning: An Approach without Crystal Structure Information. *ACS Omega* **2021**, *6* (22), 14533-14541.
22. Chen, D.; Jiang, F.; Fang, L.; Zhu, Y. B.; Ye, C. C.; Liu, W. S., Machine learning assisted discovering of new M2X3-type thermoelectric materials. *Rare Metals* **2022**, *41* (5), 1543-1553.

23. Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i., Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *Acs Catalysis* **2020**, *10* (3), 2260-2297.

24. Lamoureux, P. S.; Winther, K. T.; Torres, J. A. G.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bliagaard, T., Machine Learning for Computational Heterogeneous Catalysis. *Chemcatchem* **2019**, *11* (16), 3579-3599.

25. Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C., Machine learning for heterogeneous catalyst design and discovery. *Aiche Journal* **2018**, *64* (7), 2311-2323.

26. Ng, M.-F.; Zhao, J.; Yan, Q.; Conduit, G. J.; Seh, Z. W., Predicting the state of charge and health of batteries using data-driven machine learning. *Nature Machine Intelligence* **2020**, *2* (3), 161-170.

27. Hu, X.; Li, S. E.; Yang, Y., Advanced Machine Learning Approach for Lithium-Ion Battery State Estimation in Electric Vehicles. *Ieee Transactions on Transportation Electrification* **2016**, *2* (2), 140-149.

28. Alvarez Anton, J. C.; Garcia Nieto, P. J.; Blanco Viejo, C.; Vilan Vilan, J. A., Support Vector Machines Used to Estimate the Battery State of Charge. *Ieee Transactions on Power Electronics* **2013**, *28* (12), 5919-5926.

29. Liu, B.; Yang, J.; Yang, H. L.; Ye, C. C.; Mao, Y. Q.; Wang, J. P.; Shi, S. Q.; Yang, J. H.; Zhang, W. Q., Rationalizing the interphase stability of Li|doped-Li(7)La(3)Zr(2)O(12)via automated reaction screening and machine learning. *Journal of Materials Chemistry A* **2019**, *7* (34), 19961-19969.

30. van de Burgt, Y.; Lubberman, E.; Fuller, E. J.; Keene, S. T.; Faria, G. C.; Agarwal, S.; Marinella, M. J.; Talin, A. A.; Salleo, A., A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nature Materials* **2017**, *16* (4), 414-.

31. Mahmood, A.; Wang, J.-L., Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & Environmental Science* **2021**, *14* (1), 90-105.

32. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, *15* (10), 1120-+

33. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547-555.

34. Nisius, B.; Bajorath, J., Molecular Fingerprint Recombination: Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types. *Chemmedchem* **2009**, *4* (11), 1859-1863.

35. Miyake, Y.; Saeki, A., Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J Phys Chem Lett* **2021**, *12* (51), 12391-12401.

36. Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun,
K., Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci Adv* **2019**, *5* (11), eaay4275.

37. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*.

38. Hert, J., Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *White Rose Research Online* **2004**, *22*, 3256-3266.

39. Skoda Petr, H. D., Exploration of Topological Torsion Fingerprints. *EEE International Conference on Bioinformatics and Biomedicine (BTBM) 2015*, 822-828.

40. Polton, D. J., INSTALLATION AND OPERATIONAL EXPERIENCES WITH MACCS (MOLECULAR ACCESS SYSTEM). *Online Review* **1982**, *6* (3), 235-242.

41. Sun, W. B.; Zheng, Y. J.; Zhang, Q.; Yang, K.; Chen, H. Y.; Cho, Y.; Fu, J. H.; Odunmbaku, O.; Shah, A. A.; Xiao, Z. Y.; Lu, S. R.; Chen, S. S.; Li, M.; Qin, B.; Yang, C.; Frauenheim, T.; Sun, K., Artificial Intelligence Designer for Highly-Efficient Organic Photovoltaic Materials. *J Phys Chem Lett* **2021**, *12* (36), 8847-8854.

42. Landrum, G., RDKit: Open-source cheminformatics from machine learning to chemical registration. *Abstracts of Papers of the American Chemical Society* **2019**, 258.

43. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I., From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2020**, *2* (1), 56-67.

44. Yang, Z.; Sheng, Y.; Zhu, C.; Ni, J.; Zhu, Z.; Xi, J.; Zhang, W.; Yang, J., Accurate and explainable machine learning for the power factors of diamond-like thermoelectric materials. *Journal of Materiomics* **2022**, *8* (3), 633-639.

45. Weininger, D., SMILES, A CHEMICAL LANGUAGE AND INFORMATION-SYSTEM .1. INTRODUCTION TO METHODOLOGY AND ENCODING RULES. *Journal of Chemical Information and Computer Sciences* **1988**, *28* (1), 31-36.

46. Wadsworth, A.; Chen, H.; Thorley, K. J.; Cendra, C.; Nikolka, M.; Bristow, H.; Moser, M.; Salleo, A.; Anthopoulos, T. D.; Sirringhaus, H.; McCulloch, I., Modification of Indacenodithiophene-Based Polymers and Its Impact on Charge Carrier Mobility in Organic Thin-Film Transistors. *J Am Chem Soc* **2020**, *142* (2), 652-664.

47. Paterson, A. F.; Singh, S.; Fallon, K. J.; Hodsden, T.; Han, Y.; Schroeder, B. C.; Bronstein, H.; Heeney, M.; McCulloch, I.; Anthopoulos, T. D., Recent Progress in High-Mobility Organic Transistors: A Reality Check. *Adv Mater* **2018**, e1801079.

48. Ortiz, R. P.; Yan, H.; Facchetti, A.; Marks, T., Azine- and Azole-Functionalized Oligo’ and Polyythiophene Semiconductors for Organic Thin-Film Transistors. *Materials* **2010**, *3* (3), 1533-1558.

49. Li, M.; Wang, J.; Xu, W.; Li, L.; Pisula, W.; Janssen, R. A. J.; Liu, M., Noncovalent semiconducting polymer monolayers for high-performance field-effect transistors. *Progress in Polymer Science* **2021**, *117*.

50. Li, M.; An, C.; Pisula, W.; Mullen, K., Cyclopentadithiophene-Benzothiadiazole Donor-Acceptor Polymers as
Prototypical Semiconductors for High-Performance Field-Effect Transistors. Acc Chem Res 2018, 51 (5), 1196-1205.

51. Lei, T.; Wang, J. Y.; Pei, J., Design, synthesis, and structure-property relationships of isoindigo-based conjugated polymers. Acc Chem Res 2014, 47 (4), 1117-26.

52. Kim, Y.; Lim, E., Development of Polymer Acceptors for Organic Photovoltaic Cells. Polymers 2014, 6 (2), 382-407.

53. Hu, W.; Feng, L.; Gu, P.; Dong, H.; Yao, Y., High-mobility polymeric semiconductors. Chinese Science Bulletin 2015, 60 (23), 2169-2187.

54. He, K.; Kumar, P.; Yuan, Y.; Li, Y., Wide bandgap polymer donors for high efficiency non-fullerene acceptor based organic solar cells. Materials Advances 2021, 2 (1), 115-145.

55. Facchetti, A., π-Conjugated Polymers for Organic Electronics and Photovoltaic Cell Applications. Chemistry of Materials 2010, 23 (3), 733-758.

56. Chen, H., Recent Advances in High-Mobility Polymeric Semiconductor Materials. Chinese Journal of Organic Chemistry 2016, 36 (3).

57. Brebels, J.; Manca, J. V.; Lutsen, L.; Vanderzande, D.; Maes, W., High dielectric constant conjugated materials for organic photovoltaics. Journal of Materials Chemistry A 2017, 5 (46), 24037-24050.

58. Cao, N. D., MolGAN: An implicit generative model for small molecular graphs. Arxiv 2018.

59. Breiman, L., Random forests. Mach Learn 2001, 45 (1), 5-32.

60. Patrick, E. A.; Fischer, F. P., A GENERALIZED K-NEAREST NEIGHBOR RULE. Information and Control 1970, 16 (2), 128-&.

61. Kowalski, B. R.; Bender, C. F., K-NEAREST NEIGHBOR CLASSIFICATION RULE (PATTERN-RECOGNITION) APPLIED TO NUCLEAR MAGNETIC-RESONANCE SPECTRAL INTERPRETATION. Analytical Chemistry 1972, 44 (8), 1405-&.

62. Seeger, M., Gaussian processes for machine learning. International journal of neural systems 2004, 14 (2), 69-106.

63. Murphy, K. P., Machine Learning: A Probabilistic Perspective. The MIT Press: 2012; pp 492-493.

64. Lightbody, G.; Irwin, G. W., Multi-layer perceptron based modelling of nonlinear systems. Fuzzy Sets and Systems 1996, 79 (1), 93-112.

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

66. Breiman, L., Bagging predictors. Mach Learn 1996, 24 (2), 123-140.

67. Shi, Y.-q., Recent Progress of Imide-functionalized N-type Polymer Semiconductors. ACTA POLYMERICA SINICA.

68. Lei, T.; Cao, Y.; Fan, Y.; Liu, C.-J.; Yuan, S.-C.; Pei, J., High-Performance Air-Stable Organic Field-Effect Transistors: Isoindigo-Based Conjugated Polymers. Journal of the American Chemical Society 2011, 133 (16), 6099-6101.

69. Turkoglu, G.; Cinar, M. E.; Ozturk, T., Thiophene-Based Organic Semiconductors. Topics in Current Chemistry 2017,
375 (6).

70. Lin, Y.; Fan, H.; Li, Y.; Zhan, X., Thiazole-Based Organic Semiconductors for Organic Electronics. Advanced Materials 2012, 24 (23), 3087-3106.

71. Kang, B.; Kim, H. N.; Sun, C.; Kwon, S.-K.; Cho, K.; Kim, Y.-H., pi-Extended Thiazole-Containing Polymer Semiconductor for Balanced Charge-Carrier Mobilities. Macromolecular Rapid Communications 2021, 42 (9).

72. Shi, Q.; Zhang, S.; Zhang, J.; Oswald, V. F.; Amassian, A.; Marder, S. R.; Blakey, S. B., (KOBu)-Bu-t-Initiated Aryl C-H Iodination: A Powerful Tool for the Synthesis of High Electron Affinity Compounds. Journal of the American Chemical Society 2016, 138 (12), 3946-3949.

73. Kurata, H.; Takakuwa, H.; Imai, N.; Matsumoto, K.; Kawase, T.; Oda, M., Synthesis and properties of a 1,3-thiazole extended p-Terphenoquinone. A new sp(2) nitrogen atom-containing terquinone derivative. Bulletin of the Chemical Society of Japan 2007, 80 (7), 1402-1404.

74. Ie, Y.; Sato, C.; Yamamoto, K.; Nitani, M.; Aso, Y., A Thiazole-fused Antiaromatic Compound Containing an s-Indacene Chromophore with a High Electron Affinity. Chemistry Letters 2018, 47 (12), 1534-1537.

75. Chavez, P.; Bulut, I.; Fall, S.; Ibraikulov, O. A.; Chochos, C. L.; Bartringer, J.; Heiser, T.; Leveque, P.; Leclerc, N., An Electron-Transporting Thiazole-Based Polymer Synthesized Through Direct (Hetero)Arylation Polymerization. Molecules 2018, 23 (6).

76. Wu, N.; He, Z.-Q.; Xu, M.; Xiao, W.-K., Recent Developments of Azatriphenylene Materials as n-Type Organic Semiconductors. Acta Physico-Chimica Sinica 2014, 30 (6), 1001-1016.

77. Min, Y.; Dou, C.; Tian, H.; Geng, Y.; Liu, J.; Wang, L., n-Type Azaacenes Containing B <- N Units. Angewandte Chemie-International Edition 2018, 57 (7), 2000-2004.

78. Chai, J.; Wang, C.; Jia, L.; Pang, Y.; Graham, M.; Cheng, S. Z. D., Synthesis and electrochemical properties of a new class of boron-containing n-type conjugated polymers. Synthetic Metals 2009, 159 (14), 1443-1449.

79. Zheng, X.; Huang, R.; Zhong, C.; Xie, G.; Ning, W.; Huang, M.; Ni, F.; Dias, F. B.; Yang, C., Achieving 21% External Quantum Efficiency for Nondoped Solution-Processed Sky-Blue Thermally Activated Delayed Fluorescence OLEDs by Means of Multi-(Donor/Acceptor) Emitter with Through-Space-/Bond Charge Transfer. Advanced Science 2020, 7 (7).

80. Yang, Y.; Chen, W.; Dou, L.; Chang, W.-H.; Duan, H.-S.; Bob, B.; Li, G.; Yang, Y., High-performance multiple-donor bulk heterojunction solar cells. Nature Photonics 2015, 9 (3), 190-198.

81. Shin, S.-C.; Vincent, P.; Bae, J.-H.; Lee, J. J.; Nam, M.; Ko, D.-H.; Kim, H.; Shim, J. W., Quaternary indoor organic photovoltaic device demonstrating panchromatic absorption and power conversion efficiency of 10%. Dyes and Pigments 2019, 163, 48-54.

82. Mohammadi, E.; Kafle, P.; Huang, K.-Y.; Zhu, W.; Huang, J.; Jung, S.-H.; Lee, J.-K.; Evans, C. M.; Diao, Y., Role of Multivalent Interactions in Dynamic-Template-Directed Assembly of Conjugated Polymers. Acs Applied Materials & Interfaces 2020, 12 (2), 2753-2762.
1. Polymer-units smiles

(1) Bratch-chain

11.'CCCCCCCC(CC)CCCCC',
27.'CCCCCCCCCCCCCCCCCCCCCCCCC',
29.'CCCCCCCCCCCCCC(CC)CCCCCCCCC',
31.'CCCCCCCCCCCCC(CC)CCCCCCCCC',
39.'CCCCCCCCCCCCC(C)CCCCCCCC',
42.'CCCCCCCCCCCCC(CC)CCCCCCCCC',
46.'CCCCCCCCCCCCC(CC)CCCCCCCCC',
51.'CCCCCCCCCCCCC(CC)CCCCCCCCC',
56.'CCCCCCCCCCCCC(CC)CCCCCCCC',
62.'CCCCC(C)CC',
65.'CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC',
68.'CCCCCCCCC(C)CCCCCCCC',
69.'CCCCCCCCC',
75.'CCCCCCCCCCCCCCCCCCCCCCCCC',
87.'CCCCCCCCCCCCCCCCCCCCCCCCC',
109.'CCCCCCCCCCCCCCCCCCCCCCCCCCCCC',
114.'CCCCCCCCCCCCCCCCCCCCCCCCCCCCC',
144.'CCC(C)CCCC(C)C',
147.'CCCCCCCCCCCCCCCCCCCCC',
158.'CCCCCCCCCCCCCCCCCCCCCCCCC',
163.'CCCCCCCCCCCCCCCCCCCCCCCCC',
167.'CCCCCCCCCCCCCCCCCCCCCCCCC',
171.'CCCCCCCCCCCCCCCCCCCCCCCCC',
181.'CCCCCCCCCCCCCCCCCCCCCCCCC',
191.'CCCCCCCCCCCCCCCCCCCCCCCCC',
193.'CCCCCCCCCCCCCCCCCCCCCCCCC',
195.'CCCCCCCCCCCCCCCCCCCCCCCCC',
196.'CCCCCCCCCCCCCCCCCCCCCCCCC',
206.'CCCCCCCCCCCCCCCCCCCCCCCCC',
213.'CCCCCCCCCCCCCCCCCCCCCCCCC',
214.'CCCCCCCCCCCCCCCCCCCCCCCCC',
216.'CCCCCCCCCCCCCCCCCCCCCCCCC',
221.'CCCCCCCCCCCCCCCCCCCCCCCCC',
231.'CCCCCCCCCCCCCCCCCCCCCCCCC',
238.'CCCCCCCCCCCCCCCCCCCCCCCCC',
250.'CCCCCCCCC',
252.'CCCCCCCCC',
257.'CCCCC',
258.'CCCCC',
265.'CCCCCCCCC=CCCCCCCCC',
267.'CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC',
270.'CCCCCC(CC)CCC',
276.'CCCCCCCCCC(CC)CCCCC',
283.'CCCCCCCCCCCCCCCCCCCCC',
285.'CCCCCCCCCCCCCCCCCCCCC',
(2) Chain-F
94.'CCCCCCCCC(F)(F)(F)(F)(F)(F)(F)(F)F',
253.'FC(F)F',
287.'CCCCCCCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
293.'CCCCCCCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
318.'CCCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
370.'CCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
387.'CCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
435.'CCCCCCCCCCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
(3) Chain-F-O
197.'CCCCCOCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
294.'CCCCCOCCC(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)(F)F',
(4) Chain-N
129.'CCC#N',
146.'CCCCCCCCCCCCCN',
159.'CCCCCCCCCCN',

248. 'C=C(C)C#N',
305. 'CC=CC#N',
311. 'N#CCC#N',
313. 'C=C(C#N)C#N
(5) Chain-N-F
43. 'CCCCCCCCCCCCCCC(CCCCCCCC)CNB(F)F'
369. 'NB(F)F'
(6) Chain-O
2. 'CCCCCCCCCCCCCCCCC(CCCCCCCCCCCCCCCCC)OC(=O)CCC',
6. 'CCCCCCC(CCCC)OC=O',
9. 'CCCCCCCCCOO',
15. 'CCOCCOCO',
21. 'COCCOCO',
23. 'CCCCCCCCCOCCOCCOCCO',
25. 'CCCCCCCCCCCCOCC(=O)CCC',
26. 'CCCCCCC(CO)CCC',
55. 'CCCCCCCCCOCOCOCCOCCOCCO',
57. 'CCCCCCCCCCCCO',
59. 'CC=O',
71. 'CCCCCCC(O)CCCC',
74. 'COCCOCCOCOCO',
77. 'CCCCCCCCCO',
84. 'CCOCCOCOCO',
96. 'CCCCCCCCCCCCCCCCOC(=O)CCCCC',
97. 'CCOCCOCOCO',
105. 'CCCCCCCCCCCCCCC(OCCOCC)CCCCCCCC',
111. 'C#CO(C)CO',
113. 'CCCCCCCCCCCCCOC(=O)CCCCC',
120. 'CCCCCCCOCOCOCCOCCO',
125. 'CCCCCCCCCCCCCOCCO',
134. 'CCCCCCCCCCCCCOC(=O)CCC',
139. 'CCCCCCCCCCCC=O',
157. 'CCCCCCCCCCCCCCCCOC=O',
166. 'CCCCCCCCCOCCOCCOCCO',
174. 'CCCCCCCCCCCCC(C=O)CCCCCCCCCCCC',
192. 'CCCCCCCCCCCCCCCCCC(CCO)CCCCCCCCCCC',
208. 'CCCCC(CC)COC=O',
215. 'COCCOCCOCCOCO',
217. 'CCCCCCCOCOC=O',
225. 'CCCCCCCCCCCC(O)CCCCCCCC',
227. 'CC(C)CCC(C)COC=O',
233. 'CCCCOCO',
241. 'CCCCCCCCCCCCCCCCC(CCO)CCCCCCCCCCC',
255. 'CCOCCOCOCO',
260. 'CCCCCCCCCCCCCCCCO',
261. 'CCCCCCCCCCCCCCCCC(CCOCC)COC(=O)CCC',
282. 'CCCCCCCCCCCCC(OCCOCC)COC(=O)CCC',
292. 'CCCCCCCCCCCC(CCCCCCC)OC(=O)CCC',
295. 'CCCCCCCCCCCCCCCCCC(CO)CCCCCCCCC',
19. C(Si)(C)(C)O[Si](C)(C)O[Si](C)(C)C
73. CCCCCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
93. CCCCCCCCCC[Si](O)OC
118. CCCCCC[Si](O)[Si](C)(C)C]O[Si](C)(C)C
119. CCCCCC[Si]O[Si]C
154. C(Si)(C)(C)O
173. CCCCCCCCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
244. CCCCCCCCCC[Si]O[Si]C
246. CCCCCCCCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
264. CCCCC[Si]O[Si]C
291. CCCCCC[Si]O[Si]C
297. CCCCC[Si](C)O[Si](C)(C)C]O[Si](C)(C)C
326. CCCCCC[Si](C)(O[SiH](C)C)O[Si](C)(C)C
349. CCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
365. CCCC[Si]O[Si]C
407. CCCCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
429. CCCCCC[Si](C)(O[Si](C)(C)C)O[Si](C)(C)C
(8) Chain-S
130. CCCCCCCS
143. CCCCCC(S)CC
420. CCCCCCCCCC(S)CCCCC
(9) Chain-S-O
180. CC(=O)S
(10) Chain-Si
70. CCCCCC[Si](C)(CCCCC)CCCCC
148. CCCCCC[Si](C)(CCCCC)CCCCCCC
150. CCCCCC[Si](C)(CCCC)CCCCC
168. C#CC#C[Si](C)(CCCCC)CCCCC
219. C[Si](CC)CCC
234. CCCCC[Si]C
418.Cc1scc2c1C(=O)c1e(C)sc(C)c1C2=O
(36)Fused-S-Si
296.Cc1cc2c(s1)-c1cc3e(cc1[Si])2-c1scc1[Si]3
312.Cc1cc2e(s1)-c1scc1[Si]2
356.Cc1cc2e(s1)-c1sc(C)ce1[Si]2
378.Cc1scc2c1[Si](C)(C)c1scc1-2
(37)Fused-Si
52.Cc1cc2c([se])1-c1[cse]c(C)c1C2
223.Cc1cc2c([se])1-c1cc3e(cc1C2(C)CCCCCCCCCCCCCCCCCC)-c1[se]c(C)c1C3
(39)Mono
17.C#Cc1ccc(C)cc1
54.Cc1cc(CC)cc1
79.CC=CC1ccc(C)cc1
92.Cc1ccee(C)cc1
126.C=Cc1ccee(C)cc1
160.C#Cc1ccee(C)cc1
188.Cc1cece1C
218.CC1=CC=C(C)C1(C)C
220.e1cece1
247.Cc1cece1
286.C=CC1=CC=C(C)C
342.CC=CC1cccc1C
440.CC#Cc1ccc(C)c(C)c1
442.CC1=C(C)C(C)=C(C)CC1
(40)Mono-F
67.Cc1cece1F
72.Cc1ce(F)cc(F)c1
151.Cc1cece(F)cc1F
179.C=Cc1cece(C)cc1F
194.Cc1e(F)cc(F)cc(F)c1F
251.Cc1cece(F)c1
273.Cc1cece(C=C)cc1F
303.Cc1e(F)ccce1F
412.Cc1ce(F)ccce1F
443.Cc1cecc(F)c1
(41)Mono-N
8.Cc1cenecn1
30.Ce1cen(C#N)cc1C#N
38.Cc1cen(C)e(C)c1
80.CNc1cen(C)cc1
86.Cc1cen(C)nc1
359.CNc1cen(C)cc1C
360.Cc1ccen1
383.Cc1cen(C)cn1
(42)mono-N-O
103.CC1=CC(C)(=O)N1C
245.CC1=CC(C)(=O)N1
43) Mono-O
45.Cc1ccc(C)oc1
81.CC1CCC(C)O1
164.Cc1ccc(=O)oc1
202.COc1ccc(OC)c(C)c1
254.Cc1cc(C)c(C)oc1
302.CC=CC1ccc(C)e(OC)oc1
358.Cc1ccco1
(44) Mono-S
0.CC1CCC(C)S1
32.Cc1cc(CCCC)c(C)s1
36.Cc1ccsc(C)e1
135.Cc1cc(C)e(C)s1
137.Cc1ccsc1C
172.Cc1cccs1
178.Cc1ccsc1C
210.Cc1ccc(C)s1
235.Cc1cc(C=C)s1
344.C=Cc1ccsc(C)s1
353.C=Cc1ccsc(C)s1
392.Cc1cc(C=C)oc1
408.CC=Cc1ccc(C)s1
(45) Mono-S-Cl
259.Cc1cc(Cl)c(s1
381.Cc1cc(Cl)c(C)s1
402.Cc1ccsc1Cl
(46) Mono-S-F
33.C=Cc1cc(C)e1F
83.Cc1cc(F)e(C)s1
89.Cc1cc(C)e(C)e1F
187.Cc1cc(F)csl
228.Cc1ccsc1F
391.Cc1ccsc1F
394.Cc1cc(C=C)e(F)e1F
(47) Mono-S-N
35.Ce1nes(C)e1C
85.Cc1ccsc1C#N
88.Cc1necs1
132.Cc1necs1
133.Ce1ncsc1C
170.Ce1nc(C#N)e(C=C)s1
209.Cc1nesc1s1
332.CC=NN=Cc1ccsc(C)s1
415.Ce1nces1
421.Cc1sc(C)e(C)sc1C#N
(48) Mono-S-N-O
149.C=C1C(=O)N(CC)C(=S)N(CC)C1=O
(49) Mono-S-O
2. Polymer-unit types

2.1 Bratch chain combination

'+chain+chain-O+chain-O-Si',
'+chain+chain-N+chain-O+chain-O-Si',
'+chain-F+chain-O+chain-O-Si',
'+chain+chain-O-Si',
'+chain-O-Si',
'+chain+chain-O',
'+chain',
'+chain+chain-N+chain-O',
'+chain-O',
'+chain+chain-N-F+chain-O',
'+chain-O+chain-S-O',
'+chain-O-Si+chain-Si',
'+chain+chain-F',
'+chain-F-O+chain-O',
'+chain+chain-Si',
'+chain-F+chain-O-Si',
'+chain-O+chain-O-Si',
'+chain+chain-O-Si+chain-Si',
'+chain+chain-F+chain-O',
'+chain-F+chain-O',
'+chain+chain-F-O+chain-O',
'+chain-O+chain-S',
'+chain+chain-O+chain-Si',
'+chain+chain-O+chain-S'

2.2 Mono-ring combination

'+mono-se',
'+mono-S',
'+mono-N+mono-S',
'+mono-S+mono-S-F',
'+mono-S-N',
'+mono-S+mono-se',
'no_mono',
'+mono-S-F',
'+mono-F+mono-S',
'+mono-F+mono-S'.
2.3 Fused/bicyclic ring combination

'+bicyclic-S-N+fused-N-O',
'+bicyclic-N-O',
'+fused+fused-N-O',
'+bicyclic-N-O+fused+fused-S',
'+bicyclic-N-O+fused',
'+fused+fused-S-N-O',
'+fused+fused-N-F-O',
'+bicyclic-S-N-F+fused+fused-N-O',
'+fused+fused-S-N-F-O',
'+bicyclic-N-O+bicyclic-S',
'+fused-N-O',
'+bicyclic-S+fused+fused-N-O',
'+bicyclic-N-O+fused+S-N',
'+bicyclic-S-N+bicyclic-S-N-O',
'+bicyclic-N-O+bicyclic-S-N',
'+bicyclic-N-O-Cl+fused',
'+bicyclic-S+fused+fused-S-N',
'+bicyclic-S-N+fused+S-N-O',
'+bicyclic-N-O+fused+fused-N-O',
'+fused+fused-N-O+fused-S',
'+bicyclic-S-N-F+fused',
'+bicyclic-S-N-F+bicyclic-S-N-O',
'+bicyclic-S-N+fused',
'+fused',
'+bicyclic-S-O+fused',
'+fused+fused-S-N',
'+bicyclic-S-N+bicyclic-S-N-O+fused',
'+bicyclic-N-O+bicyclic-S+fused-N-O',
'+fused-S-N-O',
'fused+N-O+fused-S-N',
'fused-S-N-O-Cl',
'fused+N-O+fused-O',
'fused+S-O',
'fused-fused+S-N-O',
'fused+S-N-O-Cl',
'bicyclic-S-N+fused+S-S',
'bicyclic-S-N+fused+S-S',
'bicyclic-N-O+bicyclic-S+fused',
'bicyclic-N-O+bicyclic-S-N+fused',
'bicyclic+bicyclic-S-N-O+fused',
'bicyclic-S+bicyclic-S-N-O+fused',
'bicyclic-S+fused',
'bicyclic-N',
'fused+fused+S-N',
'bicyclic+S-N+fused+fused-N',
'bicyclic-N+o+bicyclic-S+fused',
'bicyclic-N-F+fused',
'bicyclic-S-N-F+fused+S-S',
'bicyclic-S+N+fused+S-Si',
'bicyclic-S-N-F+fused+S-Si',
'bicyclic-S+bicyclic-S-N-F+fused+S-S',
'bicyclic-N+fused',
'bicyclic-S+bicyclic-S-N+fused+fused-S',
'bicyclic-S-N+fused+fused-N',
'bicyclic+S-N+fused+fused-N',
'bicyclic-N+o+bicyclic-S+fused',
'bicyclic-S-N+fused+fused-N',
'bicyclic-N-F+o+bicyclic-S-fused',
'bicyclic-S-N-F+fused+S-Si+fused-Si',
'bicyclic-N+bicyclic-N-O+fused',
'bicyclic-N',
'bicyclic-N+bicyclic-N-O+bicyclic-S+fused',
'bicyclic-S+bicyclic-S-O',
'bicyclic-S-N',
'bicyclic-N-O+fused+fused-N',
'bicyclic-S+fused+fused-N-F-O',
'bicyclic-N-O+bicyclic-S+fused+fused-S-N-O',
'bicyclic-N-O+fused+fused-S-N-O',
'bicyclic+bicyclic-N-O+fused',
'bicyclic-N-O+fused+fused-S-Si',
'no_fused',
'bicyclic-S-N+fused+fused-O',
'bicyclic-S+fused+fused-S',
'bicyclic-N-O+fused-O',
'bicyclic-N-O+fused-Si',
'bicyclic-N-O+bicyclic-S-N-O+fused+fused-S-N-O',
'bicyclic-N+bicyclic-N-O',
'bicyclic-N-O+bicyclic-S-N-O+fused',
'bicyclic-S+N+fused+fused-S-O',
'bicyclic-S-F+fused+fused-S',
'fused+fused-S+fused-S-O',
'bicyclic-N-F+fused+fused-S',

'+fused+fused-S+fused-S-N',
'+fused+fused-S+fused-S-N-F',
'+bicyclic-S-N-O+fused',
'+fused+fused-S-Si'

3. Parameter Adjustment

MLP_params=[{'hidden_layer_sizes':range(50,150,10), 'activation':['identity','logistic','tanh','relu'], 'max_iter':range(100,300,10)}]
KNN_params=[{'n_neighbors':range(1,50,1), 'weights':['uniform', 'distance'], 'leaf_size':range(1,50,1)}]
gaussian_params=[{'n_restarts_optimizer':range(0,20), 'max_iter_predict':range(50,150,10), 'multi_class':['one_vs_rest','one_vs_one']}]
SVM_cl_params=[{'C': param_range_svm, 'gamma': np.logspace(-10, 3, 5)}]
RF_params=[{'max_depth': range(1,30,1), 'max_features':['auto', 'sqrt', 'log2'], 'n_estimators':range(5,200,5)}]

Complement algorithm:

SGD: SGDClassifier
LDA: LinearDiscriminantAnalysis
TREE: DecisionTreeClassifier

SGD_params=[{'loss':['hinge','modified_huber'], 'penalty':['l2','l1'], 'alpha':[i*0.0001 for i in list(range(1,5,))]}]
GDBT_params=[{'loss':['deviance','exponential'], 'n_estimators':range(5,200,5), 'criterion':['friedman_mse','squared_error','mse'], 'learning_rate':[i*0.1 for i in list(range(1,5))]}]
LDA_params=[{'solver':['svd', 'lsqr', 'eigen'], 'store_covariance':['True','False']}]
TREE_params=[{'criterion':['gini','entropy'], 'splitter':['best','random'], 'max_depth':range(1,30,1)}]
## Table S1 Cross Validation Data-P type

| CV \ Algorithm | Gaussian | KNN | MLP | RF | SVM |
|----------------|---------|-----|-----|----|-----|
| 1              | 0.704545455 | 0.75 | 0.704545455 | 0.704545455 | 0.659090909 |
| 2              | 0.659090909 | 0.704545455 | 0.681818182 | 0.681818182 | 0.590909091 |
| 3              | 0.704545455 | 0.727272727 | 0.727272727 | 0.727272727 | 0.704545455 |
| 4              | 0.727272727 | 0.704545455 | 0.681818182 | 0.681818182 | 0.795454545 |
| 5              | 0.681818182 | 0.704545455 | 0.613636364 | 0.681818182 | 0.636363636 |
| 6              | 0.772727273 | 0.681818182 | 0.727272727 | 0.818181818 | 0.7 |
| 7              | 0.681818182 | 0.659090909 | 0.681818182 | 0.727272727 | 0.590909091 |
| 8              | 0.69764419 | 0.674418605 | 0.69764419 | 0.720930233 | 0.69764419 |
| 9              | 0.69764419 | 0.627906977 | 0.674418605 | 0.651162791 | 0.674418605 |
| 10             | 0.744186047 | 0.790697674 | 0.88372093 | 0.790697674 | 0.790697674 |
| mean           | 0.707135307 | 0.702484144 | 0.716490486 | 0.723097252 | 0.689006342 |

| CV \ Algorithm | SGD | GDBT | LDA | Tree |
|----------------|-----|------|-----|------|
| 1              | 0.636363636 | 0.636363636 | 0.568181818 | 0.659090909 |
| 2              | 0.659090909 | 0.75 | 0.659090909 | 0.704545455 |
| 3              | 0.568181818 | 0.704545455 | 0.681818182 | 0.704545455 |
| 4              | 0.613636364 | 0.590909091 | 0.613636364 | 0.613636364 |
| 5              | 0.545454545 | 0.659090909 | 0.590909091 | 0.772727273 |
| 6              | 0.681818182 | 0.727272727 | 0.659090909 | 0.613636364 |
| 7              | 0.681818182 | 0.704545455 | 0.613636364 | 0.704545455 |
| 8              | 0.651162791 | 0.674418605 | 0.604651163 | 0.69764419 |
| 9              | 0.744186047 | 0.790697674 | 0.674418605 | 0.76744186 |
| 10             | 0.720930233 | 0.76744186 | 0.581395349 | 0.744186047 |
| mean           | 0.650264271 | 0.700528541 | 0.624682875 | 0.69820296 |
### Table S2 Cross Validation Data - N type

| CV\Algorithm | Gaussian | KNN | MLP | RF | SVM |
|--------------|----------|-----|-----|----|-----|
| 1            | 0.772727273 | 0.727272727 | 0.818181818 | 0.772727273 | 0.681818182 |
| 2            | 0.681818182 | 0.727272727 | 0.727272727 | 0.727272727 | 0.681818182 |
| 3            | 0.681818182 | 0.727272727 | 0.727272727 | 0.727272727 | 0.681818182 |
| 4            | 0.681818182 | 0.727272727 | 0.727272727 | 0.727272727 | 0.681818182 |
| 5            | 0.772727273 | 0.727272727 | 0.727272727 | 0.863636364 | 0.818181818 |
| 6            | 0.590909091 | 0.636363636 | 0.727272727 | 0.590909091 | 0.727272727 |
| 7            | 0.590909091 | 0.636363636 | 0.727272727 | 0.818181818 | 0.727272727 |
| 8            | 0.761904762 | 0.666666667 | 0.761904762 | 0.80952381 | 0.761904762 |
| 9            | 0.80952381 | 0.952380952 | 0.714285714 | 0.714285714 | 0.761904762 |
| 10           | 0.666666667 | 0.761904762 | 0.714285714 | 0.761904762 | 0.714285714 |

| Algorithm    | SGD       | GDBT      | LDA       | Tree       |
|--------------|-----------|-----------|-----------|------------|
| 1            | 0.636363636 | 0.727272727 | 0.5       | 0.772727273 |
| 2            | 0.545454545 | 0.727272727 | 0.409090909 | 0.681818182 |
| 3            | 0.636363636 | 0.727272727 | 0.590909091 | 0.590909091 |
| 4            | 0.863636364 | 0.863636364 | 0.727272727 | 0.863636364 |
| 5            | 0.636363636 | 0.681818182 | 0.590909091 | 0.818181818 |
| 6            | 0.818181818 | 0.727272727 | 0.545454545 | 0.863636364 |
| 7            | 0.727272727 | 0.636363636 | 0.545454545 | 0.818181818 |
| 8            | 0.545454545 | 0.590909091 | 0.5       | 0.5 |
| 9            | 0.666666667 | 0.666666667 | 0.571428571 | 0.666666667 |
| 10           | 0.714285714 | 0.761904762 | 0.523809524 | 0.714285714 |

| Algorithm    | SGD       | GDBT      | LDA       | Tree       |
|--------------|-----------|-----------|-----------|------------|

| Algorithm    | SGD       | GDBT      | LDA       | Tree       |
|--------------|-----------|-----------|-----------|------------|

### Table S3 Cross Validation Data - N/P type

| Algorithm    | Gaussian | KNN | MLP | RF | SVM |
|--------------|----------|-----|-----|----|-----|
| 1            | 0.781818182 | 0.854545455 | 0.818181818 | 0.763636364 | 0.818181818 |
| 2            | 0.818181818 | 0.8 | 0.872727273 | 0.872727273 | 0.836363636 |
| 3            | 0.8 | 0.763636364 | 0.872727273 | 0.854545455 | 0.872727273 |
| 4            | 0.722222222 | 0.759259259 | 0.87037037 | 0.851851852 | 0.796296296 |
| 5            | 0.777777778 | 0.796296296 | 0.814814815 | 0.833333333 | 0.796296296 |
| 6            | 0.87037037 | 0.907407407 | 0.87037037 | 0.925925926 | 0.888888889 |
| 7            | 0.833333333 | 0.777777778 | 0.87037037 | 0.851851852 | 0.833333333 |
| 8            | 0.851851852 | 0.87037037 | 0.888888889 | 0.851851852 | 0.851851852 |
|   | SGD     | GDBT    | LDA     | Tree    |
|---|---------|---------|---------|---------|
|  1 | 0.757557558 | 0.878787879 | 0.757575758 | 0.939393939 |
|  2 | 0.666666667 | 0.878787879 | 0.727272727 | 0.848484848 |
|  3 | 0.909090909 | 0.848484848 | 0.666666667 | 0.818181818 |
|  4 | 0.606060606 | 0.909090909 | 0.666666667 | 0.848484848 |
|  5 | 0.909090909 | 0.787878788 | 0.757575758 | 0.757575758 |
|  6 | 0.696969697 | 0.727272727 | 0.787878788 | 0.757575758 |
|  7 | 0.787878788 | 0.787878788 | 0.696969697 | 0.727272727 |
|  8 | 0.787878788 | 0.757575758 | 0.636363636 | 0.757575758 |
|  9 | 0.78125  | 0.71875  | 0.625    | 0.75     |
| 10 | 0.75     | 0.875    | 0.78125  | 0.875    |
| mean | 0.765246212 | 0.816950758 | 0.713352273 | 0.807954545 |
4. data supplement

Fig. S1 Other machine learning methods

Fig. S2 Distribution of data
Given that the data distribution does affect the prediction results, we randomly sampled 205 out of the 473 P-type data to make the amount of N-type and P-type data equal. Machine learning prediction of N/P classification was made in this dataset. The accuracy slipped slightly, but the maximum was still above 80% (Fig. S3).

![Fig. S3 N/P classify after data adjustment](image)

(a) Method 1: The classification method used in the manuscript—The mobility is classified as high (> 4 cm²·V⁻¹·s⁻¹), medium (1 to 4 cm²·V⁻¹·s⁻¹), or low (0 to 1 cm²·V⁻¹·s⁻¹).

(b) Method 2: The carrier mobility is transformed by e.q. S1:

\[ \mu \rightarrow \text{int}(\ln(\mu+1)) \]

c.q.S1

The carrier mobility is transformed by c.q.S1.
