Electronic Properties and Photocatalytic Hydrogen Evolution Rates for Alternating Conjugated Copolymers: Predictions and Insights by Data-Driven Models

Yuzhi Xu [& a)], *Cheng-Wei Ju* [& b)]*, *Bo Li* [c)], *Qiu-Shi Ma* [d)], *Lianjie Zhang* [a)]*, *Junwu Chen* [a)]

a) Institute of Polymer Optoelectronic Materials and Devices, State Key Laboratory of Luminescent Materials and Devices, College of Materials Science and Engineering, South China University of Technology, Guangzhou 510640, P. R. China
b) College of Chemistry, Nankai University, Tianjin 300071, China.
c) Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China.
d) School of Resource and Environmental Engineering, Hefei University of Technology, Hefei 230009, Anhui Province, China.

&These authors have contributed equally to this work and should be considered co-first authors.
* Corresponding: Cheng-Wei Ju (E-mail: nkuchemjcw@mail.nankai.edu.cn)
Lianjie Zhang (E-mail: lianjiezhang@scut.edu.cn)

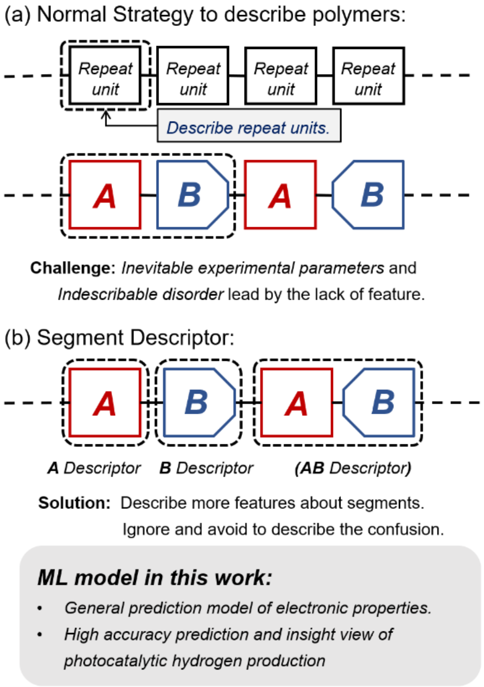
Abstract:

Alternating conjugated copolymers have been regarded as promising candidates for photocatalytic hydrogen evolution due to the adjustability of their molecular structures and electronic properties. In this work, we developed machine learning (ML) models with segment descriptors (SD) to promote the accurate and universal prediction of electronic properties without any experimental values and then constructed a high-performance prediction classifier model toward photocatalytic hydrogen production of alternating copolymers with high accuracy (real-test accuracy = 0.91). Moreover, photocatalytic dynamic study has been performed as well. Consequently, our work reveals accurate regression and classification models to disclose valuable influencing factors concerning hydrogen evolution rate (HER) of alternating copolymers.

Introduction:

Alternating conjugated copolymers incorporating different electronic units have attracted considerable attention over a wide range of opto-electronic and energy transformation applications, such as polymer light-emitting diodes[1-4], organic solar cells[5-8], organic field-effect transistors[9-11], photocatalytic hydrogen production[12-16]. Many efforts have been devoted to understanding fundamental electronic properties of alternating conjugated copolymers, such as ionization potential (IP), electron affinity (EA), optical bandgap[17-19]. In general, high hydrogen evolution rate not only relates with the optimized electronic structure but also correlates with the favored kinetic process of polymeric photocatalysts[20, 21]. From the view of the thermodynamically process, a huge body of research focus on the optimization of the electronic structure for hydrogen evolution in the last decades[22-24]. However, the study concerning the photocatalytic dynamics is scarce.

To avoid the tedious and iterative synthesis-characterization to explore the suitable electronic properties, numerous theoretical calculation methods comprising the density functional theory (DFT) have been adopted to predict the basic polymer properties approaching the experimental parameters[25, 26]. However, it is very challenging to achieve desirable predictions rapidly and accurately. Statistical learning method, as a wonderful approach, has been explored to investigate the structure-property relationship of copolymer and then seek for the suitable material candidates in organic solar cells[27-30]. However, it should be noted that such data-driven method has not been addressed well for the prediction model of polymeric hydrogen photocatalysts. Very recently, Cooper et al. navigated the available structure-property space via the integration of robotic experimentation and electronic properties of high-throughput computation[31]. Nonetheless, the experimentally measured light transmittance was needed to enhance the correlation with hydrogen evolution rate (HER). In a report by Nagasawa et al., a huge gap between the predicted PCE (5%) and experimental measurement (0.5%) was assigned to the introduction of experimental parameters and lack of description of non-order structure[32]. Towards high-throughput computation without any experimental data, advanced machine learning is deemed the desirable approach. In literature, it is still rare models combining thermodynamics with kinetics, mainly owing to the difficulty of analyzing the kinetic process by studying the stationary electronic properties. Therefore, it is of high demand to provide a method to demonstrate the relationship between ground-state electronic properties with the excited dynamic process, obtaining timely insight to give the experimenters guidance of further designing high-performance candidates.


(a) Normal Strategy to describe polymers:
Repeat unit
Repeat unit
Repeat unit
Repeat unit
Describe repeat units.
A
B
A
B
Challenge: Inevitable experimental parameters and Indescribable disorder lead by the lack of feature.
(b) Segment Descriptor:
A
B
A
B
A Descriptor
B Descriptor
(AB Descriptor)
Solution: Describe more features about segments.
Ignore and avoid to describe the confusion.
ML model in this work:
General prediction model of electronic properties.
High accuracy prediction and insight view of photocatalytic hydrogen production

Scheme 1. The motivation for Segment Descriptor.

Herein, a new class of segment descriptor (SD) inspired by the divide and conquer algorithm is proposed to describe the smallest segment of A-B alternating copolymers, which can avoid the confusion of the repeating unit's directionality (Scheme 1). Our approach can be directly applied to A-B alternating polymers even if the disorder connection of repeated units (isomers) exists. Firstly, we put forward a ML model to rapidly predict electronic properties with structure-based SD, which possesses a fantastic generalization ability and can be a wonderful tool contributing to practical research due to the impressive accuracy. Moreover, our approach shows a higher correlation in the prediction of hydrogen evolution rate (HER) compared to the basic electronic-property strategy. Based on the ML model, we adopted five classification models of high accuracy to predict the performance of polymeric hydrogen photocatalysts (testing set accuracy = 0.8). In addition, we used the testing set from different works to ensure the practicability of our model, which shows an excellent result (accuracy = 0.91). Furthermore, the relationship between the structure-property and high HER was first time to predict the design of high-performance hydrogen photocatalytic materials. 10 most crucial features are proposed thanks to the machine-selecting feature. With the decision tree classification models, relevant guidance has been demonstrated. Furthermore, an integrated insight into the mechanism in kinetic view has been supported. Delocalized excitons in excited states have been regarded as an important factor, which has been connected with electronic properties based on ML method, providing a new perspective to promote HER performance. A virtual library of the co-polymer has been generated and various building block has been ranked with our model, provided designing guidance and further demonstrated the usability of our model. Our approach provides a new strategy in facing co-polymers, proposes novel insights based on ML model, guides the scientists in further developing of high-performance HER materials.

Experiments and results:

Prediction of Electronic Properties based on Structure-based Segment Descriptor.

To resolve the disordered structure and gain more information form the copolymer, a fantastic strategy, segment descriptor (SD), is brought forward in this work. In the previous work about the ML, the alternating copolymer often adopted the A-B units as input. However, this method seems to gain limited information from the copolymer units, leading to the poor generality for further virtual screening. In contrast, SD have been regarded as a potential strategy to face the confusion and missing features. In order to prove the feasibility of segment descriptor, we first attempt to construct a rapid prediction model for the electronic properties of the alternation copolymers. To achieve quantum-mechanical-free electronic properties ML models of copolymers, molecular fingerprints, which can be inferred directly from structures of polymer units, is one of the best candidate description methods (Figure S1a) and have been widely applied in ML-based virtual screening[33]. The molecular fingerprint has proven feasible for homopolymers. Regretfully, alternating polymers cannot be simply represented in this way (only describe A-B units) due to the confusion of non-order structure, which also decreases the generality of ML models. Therefore, structure-based SD combined with the structure of each segment to describe alternating polymers can avoid this problem and give an opportunity to fulfill the handsome prediction of electronic properties.

With a dataset containing more than 6,000 simulated copolymers (containing 9 A units and 700 B units), 20% of the independent fragments B in the data set have been applied as the test set (detailed divided method can be found in Supporting Information)[31]. With the exploration of structure descriptor combination, MACCS, one type of substructure-based structure with 166 bits length have been selected to describe segment A; while circular structure, Morgan (2048 bits), have been applied as a descriptor for segment B. The success of this combination may attribute to its suitable length, while an over-long length will lead to overfitting.

Algorithms have a large influence on the accuracy and generality of ML models. Several ML algorithms have been evaluated, including Support Vector Machine (SVM), Kernel Ridge Regression (KRR), Deep Neutral Network (DNN), k-Nearest Neighbors (k-NN), LightGBM and Gradient Boost Regression Tree (GBRT). Most of these models show acceptable results (Table S1 to S2), except for k-NN owing to the high dimension of molecular fingerprints. Due to the limitation of the database size, although DNN shows high accuracy in the validation set

(Table S1), its generalization ability is mediocre, which only shows the shockingly general results in the test set (Table S2). Note that kernel function-based models such as KRR and SVR show satisfactory results in both validation set and test set. Figure 1 shows that GBRT, a tree-based ensemble model, gives the best prediction with a wonderful performance in the face of unseen segments (test set). Therefore, we can conclude that tree-based algorithms and kernel function-based algorithms can make full use of the information of the copolymer, we have constructed a ML model (GBRT/MACCS_Morgan) with generalization capabilities to predict the electronic properties of the copolymers. With the result observed above, we can confirm that SD is a feasible strategy for the representation of AB alternating copolymers.

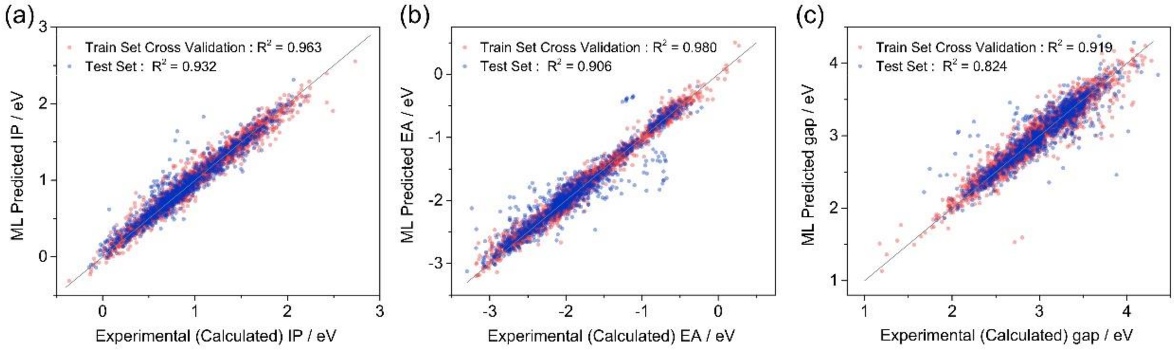


Figure 1. The linear correlation between the true (calculated) and predicted (a) IP, (b) EA and (c) gap in GBRT model with MACSS for segment A and Morgan for segment B. The red points and the blue points show the predicted result in the validation set and test set, respectively. The gray line indicates the perfect positive correlation.

Electronic Properties-based Segment Descriptor in Prediction of HER.

Furthermore, we proposed that more characteristics of the segments can increase the prediction accuracy of ML model. One important thing to note about our desired methodology is that it does not employ inputs from experiment, which allows it to be extended to a large data set and partially applied into virtual screening. Here, a valuable library containing 157 copolymers was selected because all their HER values have been reported in a previous work[31]. The dataset was randomly split into a training set (including 109 data) and test set (including 48 data). Two segments in the co-polymers (A and B) have been re-defined as Electronic Acceptor Segments (abbreviated as Acceptor) and Electronic Donor Segments (abbreviated as Donor) according to the LUMO level of the monomers (Figure S3a). Four classes of electronic properties (22 types in total) have been chosen to compose electronic properties-based SD at this time (Figure S3b, detailed information can be found in Supporting Information). Three parameters (IP, EA, Bandgap) have been adopted to describe the basic electronic properties of the co-polymer. Twelve were used to describe about the electronic structure of each segment, while another four inputs came from the difference between the energy level of two segments. The last three parameters are responsible to the dipole of the D-A unit in ground state and excited state, which can be reflected to the electronic transfer process. The Pearson correlation coefficients of the HER with these features have been displayed in Figure S4, indicating that all parameters have no significant linear correlation with HER. The main reason for such a result is due to the large influence of the experiment parameters on the performance of the copolymer and the complex mechanism, which cannot link to a single factor.

Table 1. Prediction result with different descriptors.

Descriptors	**PCCs** [a)]	**R^2**	**MAE** [b)]	**RMSE** [b)]
Electronic Properties	0.47	0.18	0.52	0.68
Segment descriptors	0.78	0.50	0.42	0.53

a) PCCs is the abbreviation of Pearson's correlation coefficient b) The unit is logarithm number.

In order to validate the stunning effectiveness of our strategy, we attempted to compare our segment descriptor (SD) method with the previous studies, where *Copper et al.* adopted several electronic properties as well as the experimental parameter into the construction of ML model, used a sub-dataset among mentioned co-polymers (fixed A unit)[31]. Based on it without any experimental data was also developed and then applied to evaluate the HER values in the used library selected from the previous work (abbreviated as *Electronic Properties*). Here, a suitable ML algorithm, Gradient Boosting Regressor, have been adopted in this comparable section. We took the logarithm of the value to gain a more reasonable distribution of the data, this is because the magnitude is more important than a real value. Table 2 and Figure S4 demonstrates the prediction result of the *Electronic Properties* and *Segment Descriptors*. We can see only employing the electronic properties (IP, EA, bandgap) as the inputs lead to a low accurate result (Pearson's correlation coefficient is only 0.47), which cannot fulfil the needs of materials prediction. Delightedly, when the *Segment Descriptors* been applied, an impressive correlation coefficient can be achieved (PCCs = 0.77). Therefore, we can conclude that since the SD contains more information, higher accuracy can be achieved than regular input, which makes it show greater potential in the application of virtual screening.

Table 2. The performance of five machine-learning classifier algorithms.

Machine-Learning techniques [a)]	Testing accuracy	Testing AUC [b)]
Extra Trees Classifier	0.80(±0.02)	0.78(±0.02)
AdaBoost Classifier	0.77	0.75
Gradient Boosting Classifier	0.77(±0.03)	0.74(±0.03)
Ridge Classifier	0.77	0.77
K-Neighbors Classifier	0.75	0.73

a) Classification accuracy was measured on the test set and training set, using the constant training dataset and the accuracy value with the standard deviation, was reported via using the average of 10 times. b) Area Under Curve of the testing set

Moreover, to achieve higher reliable prediction result as well as greater explainable, classifier model toward high-performing photocatalytic polymer was further developed. In literature a copolymer with HER higher than 1000 μmol/hg was considered as high active material candidates, which is used as a suitable judgement threshold. Five machine-learning classifier algorithms are adopted to address this problem (Table 3). All the machine-learning classifiers achieved satisfying results in test set (accuracy from 0.75 to 0.80). Notably, the Extra Trees Classifier regarded, as the most efficient one among these algorithms, obtained the highest average test accuracy (0.80), which may meet the requirement of HER prediction.

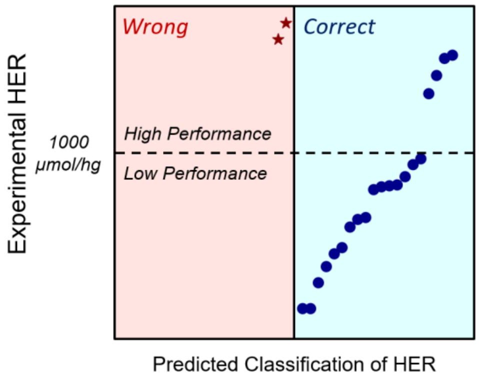


Figure 2. Prediction results versus experimental data on the external test set with the *Segment Descriptor* and Extra Tree algorithm.

Although the ML models perform well in our dataset, we expect that the HER prediction model can make sense in the real test environment instead of a toy. Therefore, 22 molecules with reasonable selecting approach (Figure S5) have been used as external test set to examine the universality of our ML model. The missing data of calculated electronic properties (IP, EA, gap) were filled with the prediction result of structure-based ML models mentioned above. To our surprise, an awesome result has been achieved (Figure 5 and Table S6), the accuracy can achieve 0.91. One of the failed examples was due to the lack of acetylene bond in our data set (8 in Figure S5), so the description of segment may be not suitable. Another failed example can be responsible to the lack of boron-embedded segment in our dataset. Although such a result may be caused by data bias probability, it can completely prove that *Segment Descriptor*-based ML model can be used in practical applications.

Hence, a high correlation coefficient model after adopting the SD method is successfully set up. After choosing from the 5 decent models, we have established an impressive high-performance HER model used in the real prediction environment with a high-accuracy result (0.91), from which we can sure our model reach a good achievement. This high-accuracy also reveals that the HER ML model has huge promising potential for application in pre-screening of hydrogen production materials and avoid the costly experiment attempt.

Insight from Machine Learning Model.

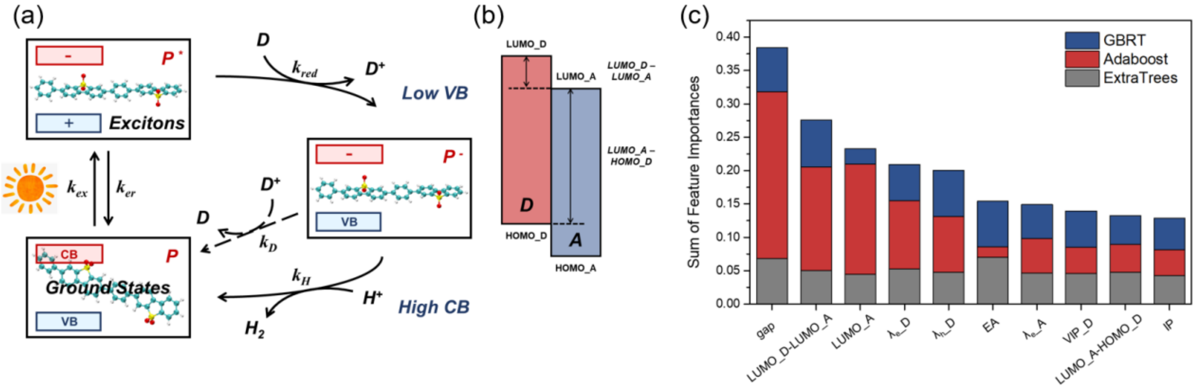


Figure 3. (a) Diagram representing the photocatalytic HER (**D** response to the sacrificial reagent). (b) Schematic of the energy level of the segments in A-B alternating co-polymers and selected energy level difference. (c) Sum of top 10 important descriptors selected by three DT-based classifier models.

Figure 3a depicts a plausible mechanism for photoinduced HER process catalyzed by conjugated copolymers. The polymer **P** absorbs light to form exciton $\mathbf{P}^*$ via electronic transition. The hole oxidizes a sacrificial reagent, leading to a stable anion $\mathbf{P}^-$ that promotes HER and, simultaneously, returns to the initial state. Note that (1) a smaller optical gap will be more suitable for molecules to be excited under blue or visible light; (2) lower VB is beneficial for the reduction by sacrificial reagent; (3) higher CB is advantageous for the HER step. Combined, as long as the polymer can be excited (i.e. the optical gap is not too big), a larger gap will be more advantageous for HER. The main side reactions are the excitons recombination process (rate is $\boldsymbol{k_{er}}$) and the reaction of **P**$^{-}$ and sacrificial positive ions (**D**$^{+}$) (rate is $\boldsymbol{k_D}$). Therefore, the rate of hydrogen production can be written as follow (detailed information can be found in Supporting Information):

$$r = \frac{k_{ex}k_{red}k_H[P][D][H]}{k_{ex}k_D[D^+]+k_{ex}k_H[H^+]+k_Dk_{red}[D][D^+]+k_Hk_{red}[D][H^+]} \qquad \text{-(1)}$$

This rate formula can provide a lot of insights and explain the characteristics of high-performance materials that were previously provided. For example, fast exciton recombination rate ($\boldsymbol{k_{er}}$) will reduce the efficiency of hydrogen production; larger surface area of the photocatalyst (**[P]**) is beneficial for hydrogen production.

It has been proposed in earlier studies that strong charge transfer is conducive to high-performance hydrogen production[34-36]. This, however, is not in line with the results of many more recent studies, where co-polymer with strong charge transfer (such as thiadiazol-embedded materials) cannot achieve high-performance (HER > 1,000 μmol/hg). Instead, many high-performance materials (such as fluorobenzene-containing copolymers) do not show strong CT features[37]. To resolve the contradictions, it is highly meaningful to develop new models and rationalize these phenomena in a more comprehensive manner. New venue opened by ML method, our approach has identified two critical descriptors, ***LUMO_D – LUMO_A*** and ***LUMO_A – HOMO_D***, that affects the HER performance of co-polymers. Both quantities describe the FMO levels of donor/acceptor segments; the former reflects the relative population of the excited electron on D/A segments, while the latter characterizes the gap of the co-polymer.

With the tree-based classifier models (GBRT, AdaBoost and Extra Trees), 10 most important features among a total of 22 ones were selected and showed in the Figure 3c. Optical band gap of the copolymer is considered as the most important feature, which is consistent to the previous works[38]. Difference between the LUMO of donor segment and acceptor segment (***LUMO_D-LUMO_A***) and the LUMO of acceptor segment (LUMO_A) are decided as the second and third most important features, respectively, implying the importance of orbital levels in photocatalysis (Figure 3b). It should not be ignored that the electron and hole reorganization energy (λ_e and λ_h) also regarded as important features, indicating that structural changes in the electron transfer process have an effect on the performance of polymers. The EA, IP of copolymers and difference between LUMO (Acceptor) with HOMO (Donor) are also considerable elements.

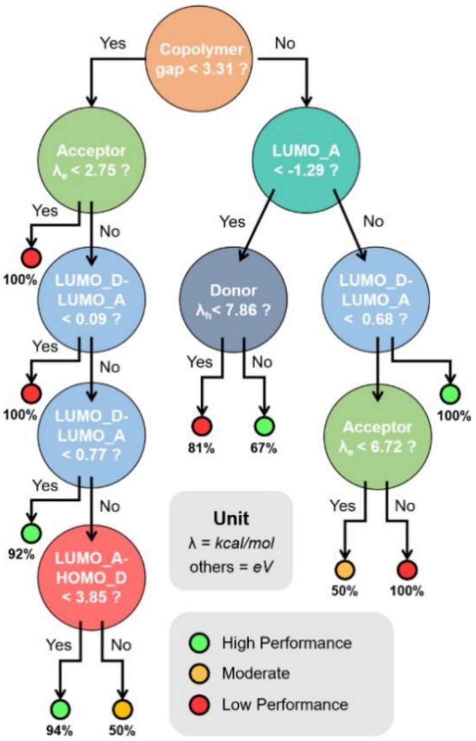


Figure 4. Best decision tree (DT) model trained with top 10 important descriptors to access different types of A-B alternating co-polymers. Categories of "High Performance" (HER>1000) and "Low Performance" (HER<1000) are colored green and red, respectively. Those that cannot be classified by a single decision tree are marked as "Moderate" in yellow color.

As noted by *Liu* et al., the fundamentals of the photocatalytic process and structure-activity relationship are still in need of more explorations[19]. To establish a relationship between electronic properties and HER performance and guide the designing of high-performance hydrogen production photocatalysts, with the Machine learning approach, a decision tree (DT) model has been constructed with the top 10 above-presented descriptors. The logical flowchart diagram of the best DT model has been shown in Figure 4. The decision tree algorithm selects the calculated bandgap as the top node in the discrimination process with a threshold of 3.31 eV. We note that the reported bandgap is calculated, and there exists a difference between experimental and calculated bandgaps; a computed 3.3 eV gap approximately corresponds to an experimental value of 2.6 eV according to the previous comparison[31]. Consistent with the previous idea, it is concluded that since the materials in the dataset are excited by blue light, the smaller gap is essential for the excitation, thus a precise range is important for further designing visible-light-excited photocatalytic hydrogen production materials[39].

The importance of charge separation/delocalization in the excited state is shown by the occurrence of $\boldsymbol{LUMO\_D - LUMO\_A}$ in several nodes in the DT. From the left side of the DT, we can notice that a $\boldsymbol{LUMO\_D - LUMO\_A}$ between 0.09 eV to 0.77 eV is more likely to cause high performance in photocatalytic hydrogen production. The low $\boldsymbol{LUMO\_D - LUMO\_A}$ typically implies a more smeared-out distribution for the excited electron. Despite the benefits of slightly concentrated orbits when a larger band gap is involved, strong CT normally brings a narrower band gap for co-polymer, which also means that high-performance materials tend to need a properly delocalized excited electron.

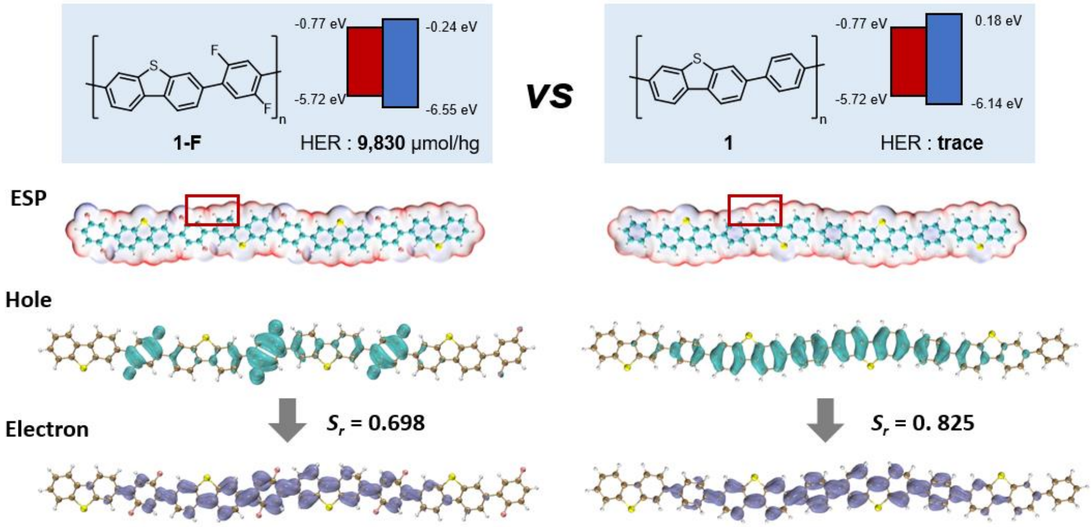


Figure 5. ESP, e-h distributions of fluorine substituted high performance co-polymer **1-F** (left) and non-fluorine substituted copolymer **1** (right) with LC-ωPBE/6-31G* (ω-tuned), green and purple represent the hole and electron distribution, respectively.

An in-depth analysis is necessary for further understanding the physical meaning of our model. A high-performance material **1-F** with HER nearly 10,000 μmol/hg, which do not show traditional electronic transfer feature, has been selected here to make (TD-)DFT calculation (Figure 5). It should note that the introduction of fluorine is extremely important, which is hundreds of times higher than **1** (trace). From the electronic static potential (ESP) of the **1** and **1-F**, negative electrostatic potential can be detected around the fluorine atom, which suggests the electron-accepting property of fluorine substituted benzene in ground state (Figure 5).

Moreover, the e-h distribution suggests that the electron in **1-F** is excited by light from a localized hole in difluorinephenyl to a more delocalized distribution over the polymer backbone, extending into dibenzothiophene.

Greater *Sr* in **1** means larger hole-electron overlap, which tends to cause higher speed in exciton recombination (i.e. excitons to ground states in Figure 3a). As a result of fast exciton recombination rate ($\boldsymbol{k_{er}}$), the excitons will recombine before the oxidation of the sacrificial agents. Thus, this can explain the low efficiency of local excitation. However, in the traditional sense, traditional strong CT tends to cause longer exciton lifetime (smaller $\boldsymbol{k_{er}}$), but why doesn't this lead to higher efficiency? As discussed in Figure 4, that a dispersed electron distribution in excited state is more favorable have been demonstrated by our model. This is because a dispersed electron distribution is believed conducive to HER. If the excited electron is localized, which is the feature or most common CT materials, reduction of H^+ will be disfavored, result in extremely small $\boldsymbol{k_H}$. Combined, it is implied that a localized hole distribution plus a delocalized electron distribution, as exemplified by **1-F**, maximize HER performance.

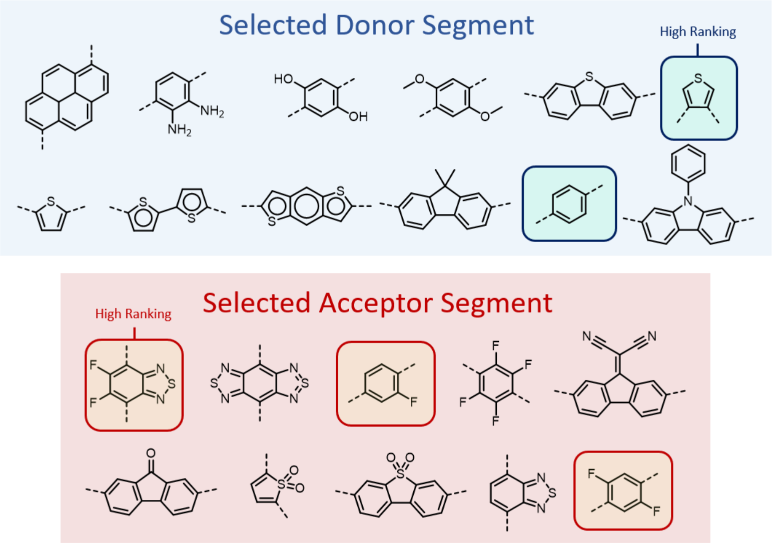


Figure 6. Selected donor segment and acceptor segment in our virtual library. High ranking segment was highlight with frame.

Based on the high accuracy of our model, we construct a virtual library with 10 commercially available donor segments and 12 common acceptor segments (Figure 6) to obtain more insights. With the HER regressor model, the ability of photocatalytic hydrogen production of each segment has been evaluated and graded based on its ranking among same class segment

(Table S7). From the result, two donor segments (1,4-phenyl and 2,3-thienyl) and three acceptor segments (1,4-difluorothiadiazole, 2-fluoro-1,4-phenyl and 2,5-difluoro-1,4-phenyl) have been recognized as more suitable candidates for high-performance hydrogen production materials (Table S8). Except for difluorothiadiazole, all these segments do not show obvious electron-poor/rich feature, which again supports the preferable lack of strong charge transfer process in most high-performance materials.

It has not escaped our notice that fluorobenzene and difluorophenyl tend to act as donors according to our models, which contrast with conventional understanding[12]. This is because the definition of donor/acceptor in our model is based on the LUMO of each segment, which can better reflect the energetic nature of the excited electron (critical to HER). The segment of fluorine substituted benzene has a low HOMO and large bandgap, which result in its electronic-poor character in ground states and potential donor part in excited states. Such a result gives a new viewpoint in supporting the high-performance of fluorine substituted photocatalyst.

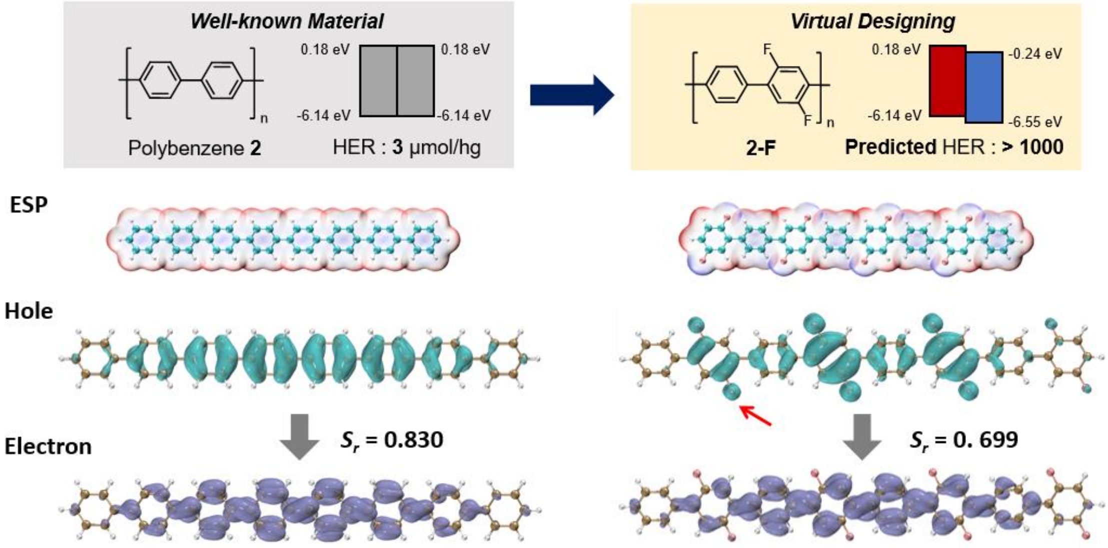


Figure 7. ESP, e-h distributions of *p*-polyphenyl **2** (left) and fluorine substituted *p*-polyphenyl **2-F** (right) with LC-ωPBE/6-31G* (ω-tuned), green and purple represent the hole and electron distribution, respectively.

Among the virtual co-polymers in our library, the optimal is fluorine *p*-polyphenyl **2-F**. Both regressor and classifier gives a prediction result with HER larger than 1,000 μmol/hg, suggest its potential high performance. (TD-)DFT analysis has been applied here to further explore its potential performance. As shown in Figure 7, with the introduction of fluorine atom, stronger charge separation is observed in **2-F**. Same as our previous viewpoint, fluorine atoms exhibit characters of donor in excited state, which suggest the better electron transfer process in excited states of **2-F**. Moreover, localized hole and delocalized electron (both in excited states) were also observed in **2-F**, illustrates its potential high performance as well.

With this model, we analyzed the 10 most important features and put forward possible mechanisms for this machine-selecting features. A logic process with details was exported to help the researchers have a better understanding of how to design high-performance materials. Insights from our model have given rise to a new viewpoint on analyzing and designing conjugated co-polymers for photocatalytic hydrogen evolution, that is, localized hole and delocalized electron are likely to promote HER performance. Further, materials optimization can be achieved by tuning the kinetic balance between decreased exciton recombination rate and increased H^+ reduction rate. Moreover, fluorine-substituted benzene can be a potential high-performance segment in designing photocatalytic hydrogen production materials co-polymers due to the exciton recombination rate can be tuned by the fluorine atom. A new fluorine substituted *p*-polyphenyl has been proposed here with predicted high-performance.

Conclusion

In summary, we have built a stunning descriptor strategy, segment descriptor, to resolve the complicated description of alternating copolymers and several novel models, including the electronic prosperities prediction model, HER regressor model, and high-performance HER classifier model. Besides, we tried to use the other molecules for other works to have a measurement. Two descriptors based on this strategy, structure-based Segment Descriptor and electronic properties-based Segment Descriptor, have been demonstrated to be feasible solutions in facing real-world problems, which provide an effective tool. This is the first time to demonstrate HER prediction model in the absence of any experimental parameter, which makes virtual high-throughput screening possible. Novel insights on discussing the importance of the co-polymer properties have been proposed. Based on the machine learning model, the dynamic analysis gives the importance of delocalized excited electron, which has been demonstrated by the further analysis of reported high-performance materials. Furthermore, based on a virtual generator, a novel co-polymer material has been proposed and proved to be potential high-performance. With the continuous studies of novel hydrogen-producing materials, more and more data will make it possible to introduce light and sacrificial agents into ML models, higher accuracy and stronger generalization capabilities will also be achieved accordingly.

Acknowledgements

We gratefully acknowledge the financial support of National Natural Science Foundation of China (51673070).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] W. Shi, S. Fan, F. Huang, W. Yang, R. Liu, Y. Cao, Synthesis of novel triphenylamine-based conjugated polyelectrolytes and their application as hole-transport layers in polymeric light-emitting diodes, J. Mater. Chem., 16 (2006) 2387-2394.

[2] B. Liu, Y.-H. Niu, W.-L. Yu, Y. Cao, W. Huang, Application of alternating fluorene and thiophene copolymers in polymer light-emitting diodes, Synth. Met., 129 (2002) 129-134.

[3] W. Ma, P.K. Iyer, X. Gong, B. Liu, D. Moses, G.C. Bazan, A.J. Heeger, Water/Methanol-Soluble Conjugated Copolymer as an Electron-Transport Layer in Polymer Light-Emitting Diodes, Adv. Mater., 17 (2005) 274-277.

[4] E. Jin, J. Li, K. Geng, Q. Jiang, H. Xu, Q. Xu, D. Jiang, Designed synthesis of stable light-emitting two-dimensional sp(2) carbon-conjugated covalent organic frameworks, Nat. Commun, 9 (2018) 4143.

[5] W. Zhong, J. Xiao, S. Sun, X.-F. Jiang, L. Lan, L. Ying, W. Yang, H.-L. Yip, F. Huang, Y. Cao, Wide bandgap dithienobenzodithiophene-based π-conjugated polymers consisting of fluorinated benzotriazole and benzothiadiazole for polymer solar cells, J. Mater. Chem. C, 4 (2016) 4719-4727.

[6] C. Liu, W. Cai, X. Guan, C. Duan, Q. Xue, L. Ying, F. Huang, Y. Cao, Synthesis of donor–acceptor copolymers based on anthracene derivatives for polymer solar cells, Polym. Chem., 4 (2013) 3949-3958.

[7] L. Meng, Y. Zhang, X. Wan, C. Li, X. Zhang, Y. Wang, X. Ke, Z. Xiao, L. Ding, R. Xia, H.L. Yip, Y. Cao, Y. Chen, Organic and solution-processed tandem solar cells with 17.3% efficiency, Science, 361 (2018) 1094-1098.

[8] Q. Liu, Y. Jiang, K. Jin, J. Qin, J. Xu, W. Li, J. Xiong, J. Liu, Z. Xiao, K. Sun, 18% Efficiency organic solar cells, Sci. Bull., 65 (2020) 272-275.

[9] P. Sista, M.P. Bhatt, A.R. Mccary, H. Nguyen, J. Hao, M.C. Biewer, M.C. Stefan, Enhancement of OFET performance of semiconducting polymers containing benzodithiophene upon surface treatment with organic silanes, J. Polym. Sci., Part A: Polym. Chem., 49 (2011) 2292-2302.

[10] C.A. Di, F. Zhang, D. Zhu, Multi-functional integration of organic field-effect transistors (OFETs): advances and perspectives, Adv. Mater., 25 (2013) 313-330.

[11] S. Allard, M. Forster, B. Souharce, H. Thiem, U. Scherf, Organic semiconductors for

solution-processable field-effect transistors (OFETs), Angew. Chem. Int. Ed., 47 (2008) 4070-4098.

[12] Y. Xu, N. Mao, C. Zhang, X. Wang, J. Zeng, Y. Chen, F. Wang, J.-X. Jiang, Rational design of donor-π-acceptor conjugated microporous polymers for photocatalytic hydrogen production, Appl. Catal., B, 228 (2018) 1-9.

[13] Z. Wang, X. Yang, T. Yang, Y. Zhao, F. Wang, Y. Chen, J.H. Zeng, C. Yan, F. Huang, J.-X. Jiang, Dibenzothiophene dioxide based conjugated microporous polymers for visible-light-driven hydrogen production, ACS Catal., 8 (2018) 8590-8596.

[14] Y. Wang, A. Vogel, M. Sachs, R.S. Sprick, L. Wilbraham, S.J.A. Moniz, R. Godin, M.A. Zwijnenburg, J.R. Durrant, A.I. Cooper, J. Tang, Current understanding and challenges of solar-driven hydrogen generation using polymeric photocatalysts, Nat. Energy, 4 (2019) 746-760.

[15] C. Dai, S. Xu, W. Liu, X. Gong, M. Panahandeh-Fard, Z. Liu, D. Zhang, C. Xue, K.P. Loh, B. Liu, Dibenzothiophene-S,S-Dioxide-Based Conjugated Polymers: Highly Efficient Photocatalyts for Hydrogen Production from Water under Visible Light, Small, 14 (2018) e1801839.

[16] E. Jin, Z. Lan, Q. Jiang, K. Geng, G. Li, X. Wang, D. Jiang, 2D sp2 carbon-conjugated covalent organic frameworks for photocatalytic hydrogen production from water, Chem, 5 (2019) 1632-1647.

[17] C.-L. Pai, C.-L. Liu, W.-C. Chen, S.A. Jenekhe, Electronic structure and properties of alternating donor–acceptor conjugated copolymers: 3,4-Ethylenedioxythiophene (EDOT) copolymers and model compounds, Polymer, 47 (2006) 699-708.

[18] P. Guiglion, A. Monti, M.A. Zwijnenburg, Validating a Density Functional Theory Approach for Predicting the Redox Potentials Associated with Charge Carriers and Excitons in Polymeric Photocatalysts, J. Phys. Chem. C, 121 (2017) 1498-1506.

[19] C. Dai, B. Liu, Conjugated polymers for visible-light-driven photocatalysis, Energy Environ.Sci., 13 (2020) 24-52.

[20] L. Wang, R. Fernández-Terán, L. Zhang, D.L. Fernandes, L. Tian, H. Chen, H. Tian, Organic Polymer Dots as Photocatalysts for Visible Light-Driven Hydrogen Generation, Angew. Chem., 128 (2016) 12494-12498.

[21] T. Hisatomi, K. Takanabe, K. Domen, Photocatalytic water-splitting reaction from

catalytic and kinetic perspectives, Catal. Lett., 145 (2015) 95-108.

[22] Y. Xu, N. Mao, S. Feng, C. Zhang, F. Wang, Y. Chen, J. Zeng, J.X. Jiang, Perylene-Containing Conjugated Microporous Polymers for Photocatalytic Hydrogen Evolution, Macromol. Chem. Phys., 218 (2017) 1700049.

[23] J. Yu, X. Sun, X. Xu, C. Zhang, X. He, Donor-acceptor type triazine-based conjugated porous polymer for visible-light-driven photocatalytic hydrogen evolution, Appl. Catal., B, 257 (2019) 117935.

[24] Y.S. Kochergin, D. Schwarz, A. Acharjya, A. Ichangi, R. Kulkarni, P. Eliášová, J. Vacek, J. Schmidt, A. Thomas, M.J. Bojdys, Exploring the “Goldilocks Zone” of semiconducting polymer photocatalysts by donor–acceptor interactions, Angew. Chem. Int. Ed., 57 (2018) 14188-14192.

[25] U. Salzner, J.B. Lagowski, P.G. Pickup, R.A. Poirier, Design of low band gap polymers employing density functional theory?hybrid functionals ameliorate band gap problem, J. Comput. Chem., 18 (1997) 1943-1953.

[26] Z. Ullah, A. Rauf, M. Yaseen, W. Hassan, M. Tariq, K. Ayub, A.A. Tahir, H. Ullah, Density functional theory and phytochemical study of 8-hydroxyisodiospyrin, J. Mol. Struct., 1095 (2015) 69-78.

[27] P.B. Jorgensen, M. Mesta, S. Shil, J.M. Garcia Lastra, K.W. Jacobsen, K.S. Thygesen, M.N. Schmidt, Machine learning-based screening of complex molecules for polymer solar cells, J. Chem. Phys., 148 (2018) 241735.

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

[29] M.-H. Lee, Robust random forest based non-fullerene organic solar cells efficiency prediction, Org. Electron., 76 (2020) 105465.

[30] D. Padula, J.D. Simpson, A. Troisi, Combining electronic and structural features in machine learning models to predict organic solar cells properties, Materials Horizons, 6 (2019) 343-349.

[31] Y. Bai, L. Wilbraham, B.J. Slater, M.A. Zwijnenburg, R.S. Sprick, A.I. Cooper, Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory, J. Am. Chem. Soc., 141 (2019) 9063-9071.

[32] S. Nagasawa, E. Al-Naamani, A. Saeki, Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest, J Phys Chem Lett, 9 (2018) 2639-2646.

[33] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, A Structure-Based Platform for Predicting Chemical Reactivity, Chem, 6 (2020) 1379-1390.

[34] W. Chen, L. Wang, D. Mo, F. He, Z. Wen, X. Wu, H. Xu, L. Chen, Modulating Benzothiadiazole-Based Covalent Organic Frameworks via Halogenation for Enhanced Photocatalytic Water Splitting, Angew. Chem. Int. Ed. Engl., 59 (2020) 16902-16909.

[35] Y.K. Eom, L. Nhon, G. Leem, B.D. Sherman, D. Wang, L. Troian-Gautier, S. Kim, J. Kim, T.J. Meyer, J.R. Reynolds, Visible-Light-Driven Photocatalytic Water Oxidation by a π-Conjugated Donor–Acceptor–Donor Chromophore/Catalyst Assembly, ACS Energy Lett., 3 (2018) 2114-2119.

[36] J. Huo, S.N. Wang, Y. Liu, X. Hu, Q. Deng, D. Chen, Arylene Ethynylene-Functionalized Bithiazole-Based Zinc Polymers for Ultraefficient Photocatalytic Activity, ACS Omega, 4 (2019) 17798-17806.

[37] B. Chen, X. Wang, W. Dong, X. Zhang, L. Rao, H. Chen, D. Huang, Y. Xiang, Enhanced Light-Driven Hydrogen-Production Activity Induced by Accelerated Interfacial Charge Transfer in Donor–Acceptor Conjugated Polymers/TiO2 Hybrid, Chem. Eur. J., 25 (2019) 3362-3368.

[38] Y. Wang, F. Silveri, M.K. Bayazit, Q. Ruan, Y. Li, J. Xie, C.R.A. Catlow, J. Tang, Bandgap Engineering of Organic Semiconductors for Highly Efficient Photocatalytic Water Splitting, Adv. Energy Mater., 8 (2018) 1801084.

[39] Y. Wang, M.K. Bayazit, S.J.A. Moniz, Q. Ruan, C.C. Lau, N. Martsinovich, J. Tang, Linker-controlled polymeric photocatalyst for highly efficient hydrogen evolution from water, Energy Environ.Sci., 10 (2017) 1643-1651.