Segment Descriptor Enabling Prediction of Electronic Properties and Photocatalytic Hydrogen Evolution Rate of Alternating Conjugated Copolymers Based on Machine Learning

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 & Devices, State Key Laboratory of Luminescent Materials & Devices, 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)

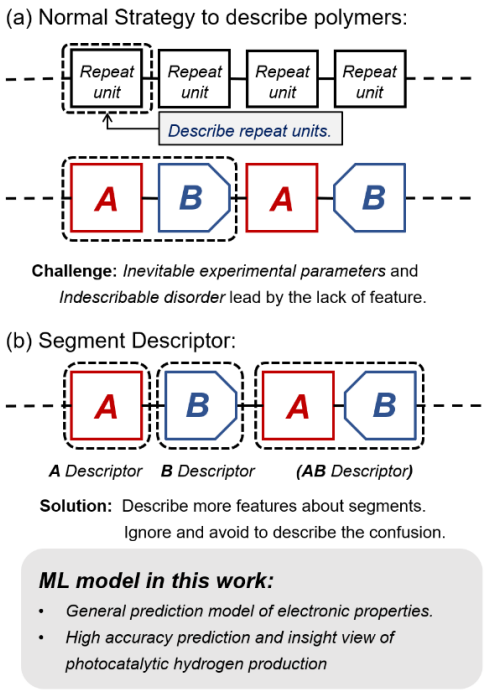
Abstract:

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, machine learning (ML) models with molecular fingerprint of segment descriptors (SD) have been successfully constructed to promote the accurate and universal prediction of electronic properties such as electron affinity, ionization potential and optical bandgap. Moreover, without any experimental values, a high-performance prediction classifier model toward photocatalytic hydrogen production of alternating copolymers has been developed with high accuracy (real-test accuracy = 0.91). Consequently, our results demonstrate 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, including ionization potential (IP), electron affinity (EA), optical bandgap[17-19]. Specifically, the optimized electronic properties are highly conductive to promote the performance of polymeric hydrogen photocatalysts[20]. To avoid the tedious and iterative synthesis-characterization, numerous theoretical calculation methods comprising the density functional theory (DFT) have been adopted to predict the basic polymer properties approaching the experimental parameters[21, 22]. However, it is very challenging to achieve desirable predictions rapidly and accurately.

Machine learning (ML) 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[23-26]. However, it should be noted that the ML method has not been addressed well for the prediction 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[27]. 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%) cannot be avoided even with the introduction of experimental parameters[28]. Towards high-throughput computation without any experimental data, advanced machine learning is deemed the desirable approach.



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 molecular fingerprints-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). Furthermore, the relationship between the structure-property and high HER was first time to predict the design of high-performance hydrogen photocatalytic materials. With the decision tree classification models, relevant guidance has been demonstrated. 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). To our best knowledge, it is the first time to combines the electronic features of the segments of copolymers in machine-learning models without experiment parameters to virtually predict the performance of hydrogen photocatalysts, which can greatly help the practicing researchers to explore the potential hydrogen photocatalytic candidates in the future. Besides, we are convinced that this segment inception can also offer a thinking path to address the more difficult problems about copolymers.

Experiments and results:

To resolve the disorder structure and gain more information from 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[28]. However, this method seems to gain the limited information from the copolymer units, leading to the poor generality for further virtual screening. In contrast, SD have been regarded 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. Molecular fingerprints have 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, fingerprints-based SD combined with the fingerprints of each segment to describe alternating polymers can avoid this problem and give an opportunity to fulfill the handsome prediction of electronic properties.

The database includes 6,354 simulated copolymers with more than 700 monomers (containing 9 A units and 700 B units), which is reported in previous work[27]. We hope to develop the ML model with the generalization capabilities of the prediction of electronic properties (IP, EA, Bandgap). Thus, the database was randomly split into two units, 80% segments of B in the database were regarded as the training set and the 20% remaining was the test set (Figure S2). In the process of adjusting the model, 50% of the data in the total training set (about 5,000 data) have been used as the validation set and the rest as the training set. The test set is only used when the model generalization ability is ultimately checked.

Exploring a suitable combination of fingerprints is important. Three types of descriptors have been proposed to express each unit at this time – Random, MACCS,

and Morgan (their lengths increase in the same order, 1 num, 166 bits, 2048 bits). In our system, the number of A unit only includes 9 monomers, which is much less than that of the B unit. Therefore, we should use longer fingerprints for B to express more features. Six combinations can be generated (Figure S1b). LightGBM (Light Gradient Boosting Machine), a tree-based algorithm featuring high-speed training, has been opted for preliminary exploration. MAE is used to evaluate the accuracy of the model, and the determination coefficient is also provided as a reference. From the result of the prediction of test set shown in Figure S1b, two conclusions can be drawn: 1) The accuracy of the prediction is mainly determined by the description of B; 2) The combination of MACCS and Morgan manifests the best accuracy, this may attribute to its suitable length, while an over-long length will lead to overfitting. Therefore, MACCS/Morgan can be regarded as the optimal fingerprint combination for this copolymer system, as it displays a satisfactory accuracy with fewer characters.

Table 1. Performance of different models measured using mean absolute error (MAE) and coefficient of determination (R^2) metrics, measured on the test set and validation set.

Machine-learning techniques	IP		EA		Bandgap	
	R^2	MAE(eV)	R^2	MAE(eV)	R^2	MAE(eV)
k-NN	0.767	0.150	0.840	0.164	0.794	0.130
SVR	0.915	0.090	0.899	0.131	0.807	0.119
KRR	0.932	0.075	0.900	0.121	0.819	0.113
LightGBM	0.932	0.073	0.880	0.123	0.822	0.111
GBRT	**0.932**	**0.073**	**0.906**	**0.118**	**0.824**	**0.108**
DNN	0.919	0.081	0.889	0.134	0.790	0.125

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 1), 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. 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.

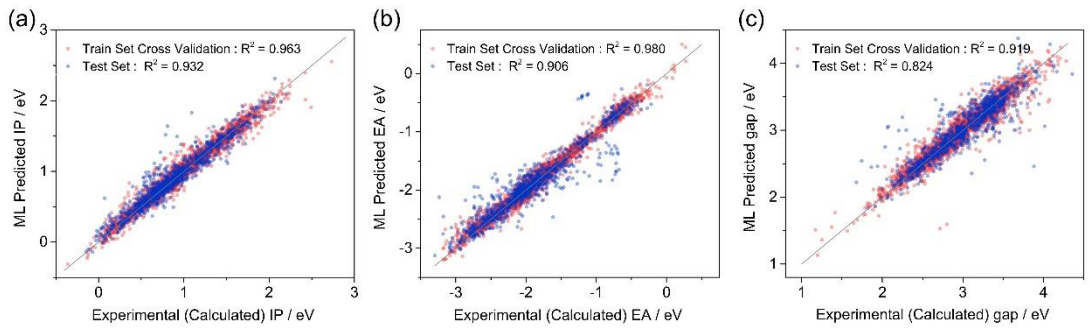


Figure 1. The linear correlation between the true (calculated) and predicted (a) IP, (b) EA and (c) gap in GBRT model with MACSS for segment A and Morgan for segment B. The red points and the blue points show the predicted result in the validation set and test set, respectively. The gray line indicates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

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[27]. The dataset was randomly split into a training set (including 109 data) and test set (including 48 data). Two segments in the copolymers

(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 2a). Four classes of electronic properties (22 types in total) have been chosen to compose electronic properties-based SD at this time (Figure 2b, detailed information can be found in Supporting Information). Three parameters (IP, EA, Bandgap) have been adopted to describe the basic electronic properties of the copolymer. Twelve were used to describe about the electronic structure of each segments, 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 demonstrated (Figure S3), indicating that all parameters have no significant linear correlation with HER. The main reason for such 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.

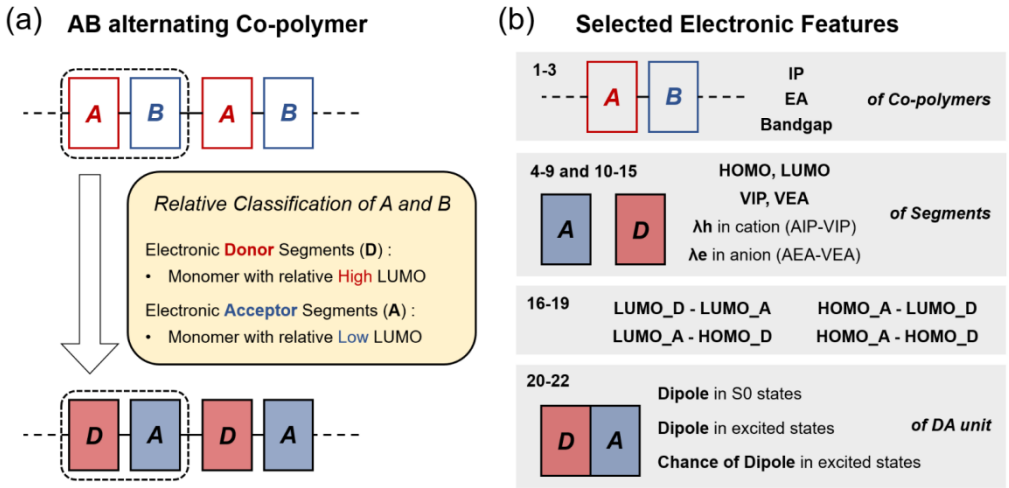


Figure 2. (a) Redefine the segment of A-B alternating copolymers into Electronic Acceptor Segments (abbreviated as Acceptor) and Electronic Donor Segments (abbreviated as Donor). (b) 22 selected microscopic properties of the redefined segment and copolymer selected as descriptors. Among these parameters, VIP means vertical ionization potential, VEA means vertical electron affinity. AIP and AEA means

adiabatic ionization potential and adiabatic electron affinity, respectively.

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 adopted several electronic properties as well as the experimental parameter into the construction of ML model, used a sub-dataset among mentioned copolymers (fixed A unit)[27]. Based on it the without any experimental data was also developed and then applied to evaluate the HER values in the used library selected from the previous work[27] (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 leads to a low accurate result (Pearson's correlation coefficient is only 0.47), which cannot fulfill 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. 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 index number.

Moreover, classifier model toward high-performing PC polymer was further developed to realize a high accurate prediction for hydrogen evolution. In literature a

copolymer with HER less than 1000 μmol/(hg) was considered as low active material candidate[27], 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.

Table 3. 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

With the tree-based classifier models (GBRT, AdaBoost and Extra Trees), 10 of the most important features among the 22 features were selected. As shown in Figure 3a, the most important feature is optical band gap of copolymer, which is consistent to the previous work[29]. Besides, difference between the LUMO of donor segment and acceptor segment (LUMO_D – LUMO_A) and the LUMO of acceptor segment (LUMO_A) working as the second and third most important features, respectively, are also observed, implying that the electronic transfer plays an important role in the 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 structural changes in the electron transfer process have effect to the performance of polymers. In addition, the EA, IP of copolymers and difference between LUMO (Acceptor) with HOMO (Donor) are also considerable elements as well.

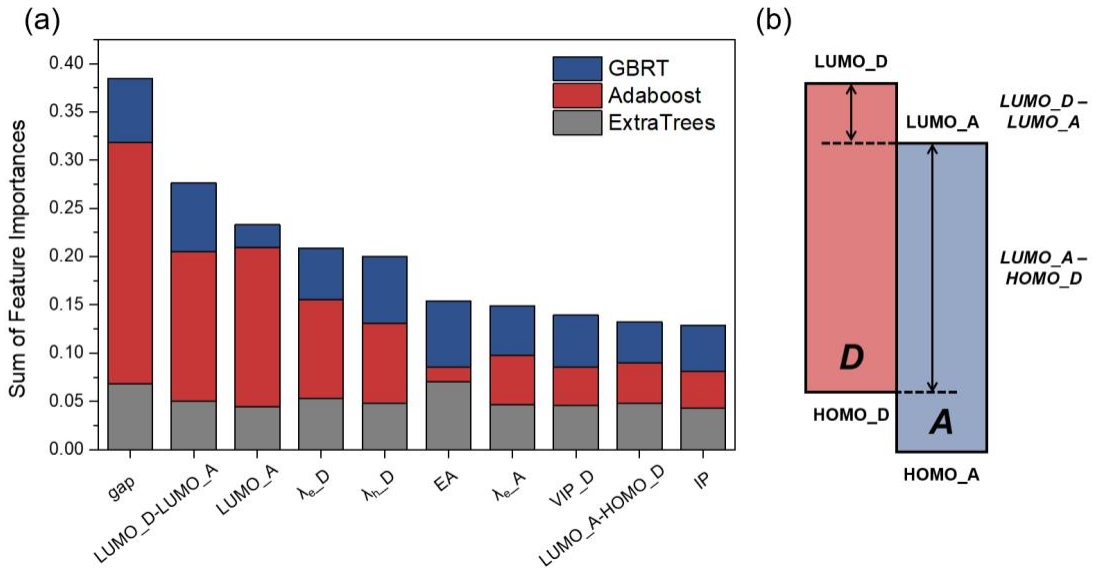


Figure 3. (a) Sum of top 10 important descriptors selected by three DT-based classifier models. (b) Schematic of the energy level of the segments in A-B alternating co-polymers and selected energy level difference.

As mentioned by Liu et al., the fundamentals of photocatalytic process and structure-activity relationship remain to have a better exploration[19]. Therefore, to explore the relationship between the electronic properties and HER performance and help us design the high-performance hydrogen production photocatalysts, a decision tree (DT) model had been constructed with top 10 important descriptors mentioned above. The logical flowchart diagram of the best DT model has been shown in Figure 4. As a result, the decision tree algorithm selects the calculated bandgap as the top node in the discrimination process with a threshold of 3.31 eV. It should be noted that the bandgap applied here is only calculated, and there is a certain difference between the experimental bandgap and the calculated bandgap. The calculated 3.3 eV approximately corresponds to the experimental value of 2.6 eV according to the previous comparison[27]. Such conclusion is consistent with the previous work, though a precise range have not been provided before[19, 30]. The process of electron flowing from LUMO of donor segment to LUMO of acceptor segment in the excited states can be demonstrated, because the difference between two energy level (LUMO_D – LUMO_A) plays as several nodes in the DT. Besides, the λ_e of the acceptor segments and the λ_h of the donor

segments have been selected by DT, which means the structure change in electronic process seems to be important, can also prove the rationality of our model.

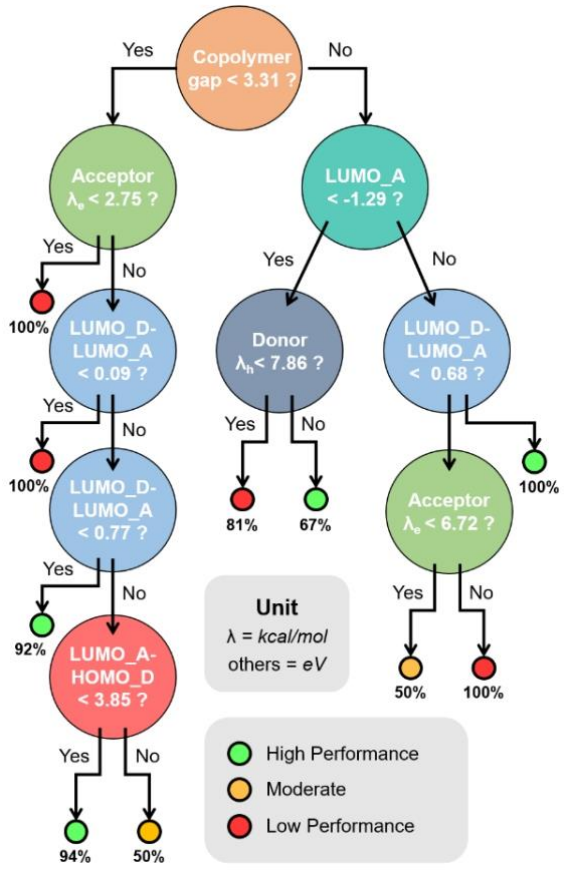


Figure 4. Best decision tree (DT) model trained with top 10 important descriptors to access different types of A-B alternating co-polymers. Categories of "High Performance" (HER>1000) and "Low Performance" (HER<1000) are colored green and red, respectively. Those that cannot be classified by a single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

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 fingerprint-based ML models mentioned above. To our surprise, an awesome result has been

achieved (Figure 5 and Table S5), the accuracy can achieve 0.91. One of the failed example was due to the lack of acetylene bond in our data set (8 in Figure S5), so 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 result may be caused by data bias probability, it can completely prove that *Segment Descriptor*-based ML model can be used in practical applications.

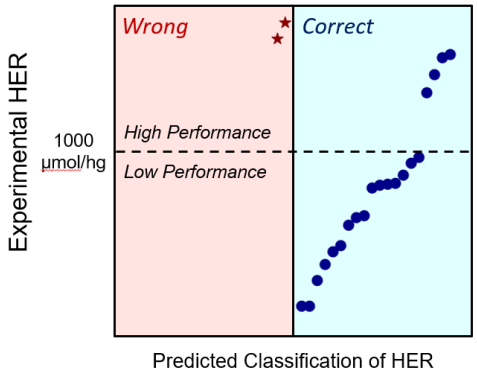


Figure 5. Prediction results versus experimental data for the predicted hydrogen photocatalysis materials with the *Segment Descriptor* and Extra Tree algorithm.

Hence, after choosing from the 5 decent model, 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. With this model, we also analyzed the 10 most important features and put forward possible mechanisms for this machine-selecting features. Besides, a logic process with details was exported to help the researchers have a better understanding of how to design high-performance materials.

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, molecular fingerprint-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. We also provide insights on discussing the importance order about the co-polymer properities and how to construct high-performance hydrogen-producing materials. 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). We sincerely thank Prof. Fei Huang in South China University of Technology for his helpful discussion in the description of hydrogen evolution part. We also thank for Ma Yu-Miao for the computer service support.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Shi, W.; Fan, S.; Huang, F.; Yang, W.; Liu, R.; Cao, Y., Synthesis of novel triphenylamine-based conjugated polyelectrolytes and their application as hole-transport layers in polymeric light-emitting diodes. *J. Mater. Chem.* **2006,** 16, (24), 2387-2394.
2. Liu, B.; Niu, Y. H.; Yu, W. L.; Cao, Y.; Huang, W., Application of alternating fluorene and thiophene copolymers in polymer light-emitting diodes. *Synth. Met.* **2002,** 129, (2), 129-134.
3. Ma, W.; Iyer, P. K.; Gong, X.; Liu, B.; Moses, D.; Bazan, G. C.; Heeger, A. J., Water/Methanol-Soluble Conjugated Copolymer as an Electron-Transport Layer

in Polymer Light-Emitting Diodes. *Adv. Mater.* **2005,** 17, (3), 274-277.

4. Jin, E.; Li, J.; Geng, K.; Jiang, Q.; Xu, H.; Xu, Q.; Jiang, D., Designed synthesis of stable light-emitting two-dimensional sp 2 carbon-conjugated covalent organic frameworks. *Nat. Commun.* **2018,** 9, (1), 1-10.
5. Wenkai, Z.; Jingyang, X.; Sheng, S.; Xiao-Fang, J.; Linfeng, L.; Lei, Y.; Wei, Y.; Hin-Lap, Y.; Fei, H.; Yong, C., Wide bandgap dithienobenzodithiophene-based p-conjugated polymers consisting of fluorinated benzotriazole and benzothiadiazole for polymer solar cells. *J. Mater. Chem. C* **2016,** 4, (21), 4719-27.
6. Liu, C.; Cai, W.; Guan, X.; Duan, C.; Xue, Q.; Ying, L.; Huang, F.; Cao, Y., Synthesis of donor-acceptor copolymers based on anthracene derivatives for polymer solar cells. *Polym. Chem.* **2013,** 4, (14), 3949-3958.
7. Cui, C.; He, Z.; Wu, Y.; Cheng, X.; Wu, H.; Li, Y.; Cao, Y.; Wong, W.-Y., High-performance polymer solar cells based on a 2D-conjugated polymer with an alkylthio side-chain. *Energy Environ.Sci.* **2016,** 9, (3), 885-891.
8. Wu, Y.; Yang, H.; Zou, Y.; Dong, Y. Y.; Yuan, J. Y.; Cui, C. H.; Li, Y. F., A new dialkylthio-substituted naphtho 2,3-c thiophene-4,9-dione based polymer donor for high-performance polymer solar cells. *Energy Environ.Sci.* **2019,** 12, (2), 675-683.
9. Sista, P.; Bhatt, M. P.; McCary, A. R.; Nguyen, H.; Hao, J.; Biewer, M. C.; Stefan, M. C., Enhancement of OFET Performance of Semiconducting Polymers Containing Benzodithiophene Upon Surface Treatment with Organic Silanes. *J. Polym. Sci., Part A: Polym. Chem* **2011,** 49, (10), 2292-2302.
10. Sung, M. J.; Kim, Y.; Lee, S. B.; Lee, G. B.; An, T. K.; Cha, H.; Kim, S. H.; Park, C. E.; Kim, Y.-H., New dithienophosphole-based donor-acceptor alternating copolymers: Synthesis and structure property relationships in OFET. *Dyes Pigm.* **2016,** 125, 316-322.
11. Bronstein, H.; Ashraf, R. S.; Kim, Y.; White, A. J. P.; Anthopoulos, T.; Song, K.; James, D.; Zhang, W.; McCulloch, I., Synthesis of a Novel Fused Thiophene-thieno 3,2-b thiophene-thiophene Donor Monomer and Co-polymer for Use in OPV and OFETs. *Macromol. Rapid Commun.* **2011,** 32, (20), 1664-1668.

12. Sprick, R. S.; Bonillo, B.; Clowes, R.; Guiglion, P.; Brownbill, N. J.; Slater, B. J.; Blanc, F.; Zwijnenburg, M. A.; Adams, D. J.; Cooper, A. I., Visible-Light-Driven Hydrogen Evolution Using Planarized Conjugated Polymer Photocatalysts (vol 55, pg 1792, 2016). *Angew. Chem. Int. Ed.* **2018,** 57, (10), 2520-2520.
13. Sachs, M.; Sprick, R. S.; Pearce, D.; Hillman, S. A. J.; Monti, A.; Guilbert, A. A. Y.; Brownbill, N. J.; Dimitrov, S.; Shi, X.; Blanc, F.; Zwijnenburg, M. A.; Nelson, J.; Durrant, J. R.; Cooper, A. I., Understanding structure-activity relationships in linear polymer photocatalysts for hydrogen evolution. *Nat. Commun.* **2018,** 9.
14. Dai, C.; Xu, S.; Liu, W.; Gong, X.; Panahandeh-Fard, M.; Liu, Z.; Zhang, D.; Xue, C.; Loh, K. P.; Liu, B., Dibenzothiophene-S,S-Dioxide-Based Conjugated Polymers: Highly Efficient Photocatalyts for Hydrogen Production from Water under Visible Light. *Small* **2018,** 14, (34).
15. Zhang, X.-H.; Wang, X.-P.; Xiao, J.; Wang, S.-Y.; Huang, D.-K.; Ding, X.; Xiang, Y.-G.; Chen, H., Synthesis of 1,4-diethynylbenzene-based conjugated polymer photocatalysts and their enhanced visible/near-infrared-light-driven hydrogen production activity. *J. Catal.* **2017,** 350, 64-71.
16. Jin, E.; Lan, Z.; Jiang, Q.; Geng, K.; Li, G.; Wang, X.; Jiang, D., 2D sp2 carbon-conjugated covalent organic frameworks for photocatalytic hydrogen production from water. *Chem* **2019,** 5, (6), 1632-1647.
17. Pai, C. L.; Liu, C. L.; Chen, W. C.; Jenekhe, S. A., Electronic structure and properties of alternating donor-acceptor conjugated copolymers: 3,4-Ethylenedioxythiophene (EDOT) copolymers and model compounds. *Polymer* **2006,** 47, (2), 699-708.
18. Guiglion, P.; Monti, A.; Zwijnenburg, M. A., Validating a Density Functional Theory Approach for Predicting the Redox Potentials Associated with Charge Carriers and Excitons in Polymeric Photocatalysts. *J. Phys. Chem. C* **2017,** 121, (3), 1498-1506.
19. Dai, C.; Liu, B., Conjugated polymers for visible-light-driven photocatalysis. *Energy Environ.Sci.* **2020,** 13, (1), 24-52.
20. Zhang, G.; Lan, Z. A.; Wang, X., Conjugated polymers: catalysts for

photocatalytic hydrogen evolution. *Angew. Chem. Int. Ed.* **2016,** 55, (51), 15712-15727.

21. Salzner, U.; Lagowski, J. B.; Pickup, P. G.; Poirier, R. A., Design of low band gap polymers employing density functional theory - Hybrid functionals ameliorate band gap problem. *J. Comput. Chem.* **1997,** 18, (15), 1943-1953.
22. Ullah, Z.; Ata ur, R.; Fazl i, S.; Rauf, A.; Yaseen, M.; Hassan, W.; Tariq, M.; Ayub, K.; Tahir, A. A.; Ullah, H., Density functional theory and phytochemical study of 8-hydroxyisodiospyrin. *J. Mol. Struct.* **2015,** 1095, 69-78.
23. Jorgensen, P. B.; Mesta, M.; Shil, S.; Lastra, J. M. G.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N., Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **2018,** 148, (24).
24. 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).
25. Lee, M.-H., Robust random forest based non-fullerene organic solar cells efficiency prediction. *Org. Electron.* **2020,** 76.
26. Padula, D.; Simpson, J. D.; Troisi, A., Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons* **2019,** 6, (2), 343-349.
27. Bai, Y.; Wilbraham, L.; Slater, B. J.; Zwijnenburg, M. A.; Sprick, R. S.; Cooper, A. I., Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. *J. Am. Chem. Soc.* **2019,** 141, (22), 9063-9071.
28. Nagasawa, S.; Al-Naamani, E.; Saeki, A., Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *J. Phys. Chem. Lett.* **2018,** 9, (10), 2639-2646.
29. Wang, Y.; Silveri, F.; Bayazit, M. K.; Ruan, Q.; Li, Y.; Xie, J.; Catlow, C. R. A.; Tang, J., Bandgap engineering of organic semiconductors for highly efficient photocatalytic water splitting. *Adv. Energy Mater.* **2018,** 8, (24), 1801084.

30. Wang, Y.; Bayazit, M. K.; Moniz, S. J. A.; Ruan, Q.; Lau, C. C.; Martsinovich, N.; Tang, J., Linker-controlled polymeric photocatalyst for highly efficient hydrogen evolution from water. *Energy Environ.Sci.* **2017,** 10, (7), 1643-1651.