ARTICLE OPEN

Data driven discovery of conjugated polyelectrolytes for optoelectronic and photocatalytic applications

Yangyang Wan¹, Fernando Ramirez¹, Xu Zhang¹, Thuc-Quyen Nguyen², Guillermo C. Bazan² and Gang Lu¹✉

Conjugated polyelectrolytes (CPEs), comprised of conjugated backbones and pendant ionic functionalities, are versatile organic materials with diverse applications. However, the myriad of possible molecular structures of CPEs render traditional, trial-and-error materials discovery strategy impractical. Here, we tackle this problem using a data-centric approach by incorporating machine learning with high-throughput first-principles calculations. We systematically examine how key materials properties depend on individual structural components of CPEs and from which the structure–property relationships are established. By means of machine learning, we uncover structural features crucial to the CPE properties, and these features are then used as descriptors in the machine learning to predict the properties of unknown CPEs. Lastly, we discover promising CPEs as hole transport materials in halide perovskite-based optoelectronic devices and as photocatalysts for water splitting. Our work could accelerate the discovery of CPEs for optoelectronic and photocatalytic applications.

npj Computational Materials (2021)7:69; <https://doi.org/10.1038/s41524-021-00541-5>

INTRODUCTION

Conjugated polyelectrolytes (CPEs) are a class of macromolecules that contain π -conjugated backbones, ionic-functionalized alkyl side chains, and counter ions^{1–4}. The delocalized π -electrons in the backbone endow CPEs with optical and charge transport properties of organic semiconductors, while the ionic functionalities render CPEs soluble in high dielectric constant media and in particular water^{4,5}. As versatile optoelectronic materials, CPEs have been used in organic and perovskite photovoltaics^{6–8}, light-emitting diodes^{9–11}, thermoelectrics¹², and photocatalysts^{13,14}. Recently, narrow bandgap CPEs (NBCPEs) have attracted particular attention as their backbones comprise electron-rich and electron-poor fragments, leading to intramolecular charge-transfer states and facile oxidation and reduction¹⁵. Remarkably, when combined with anionic sulfonate side groups, the NBCPEs could be spontaneously self-doped in aqueous media^{12,16–18}. In contrast, the self-doping is not present in cationic CPEs. This suggests that the interplay between the electrostatic force and the redox chemistry in the semiconducting backbone is key to the optoelectronic properties of NBCPEs, and particularly to stabilizing polaron states. Unfortunately, despite tantalizing potentials and significant interests in CPEs, systematic studies of CPEs remain scarce, especially from a theoretical and/or first-principles perspective. There is a critical lack of fundamental understanding on how electrostatic force may influence the optoelectronic properties of CPEs. It is unclear how structural components (backbone, side chain, ionic group, and counter ions) of CPEs may be individually and collectively tuned to optimize the optoelectronic properties. More importantly, it remains highly challenging to design CPEs for target applications.

In this work, we aim to tackle these challenges using a data-centric approach by combining machine learning with high-throughput first-principles calculations. Although the same strategy has been used in other types of materials^{19–24}, no such effort has been attempted on CPEs. Our work is further motivated by the needs to establish structure–property relationships in CPEs.

Because all structural components, including the donor and acceptor units on the backbone, the ionic identity of the pendant groups, the length of the side chains and the choice of the counter ions can be individually tuned, a large number of structural variations could emerge, rendering the development of structure–property relationships challenging. On the other hand, the structural modularity of CPEs also makes them amenable to machine learning. Thus, it is of interest to explore whether machine learning can help navigate the high-dimensional space of the structural variations and uncover much needed structure–property relationships for designing/screening CPEs. Essential to machine learning is the availability of pertinent data and in this work, we will construct a first-principles database of CPEs. In the following, we will first provide a brief description of the database and then examine how frontier orbitals, such as the highest-occupied molecular-orbital (HOMO), the lowest-unoccupied-molecular-orbital (LUMO), and the HOMO–LUMO gap (E_g) of CPEs depend on each structural component. Subsequently, by means of machine learning, we will uncover structural features of CPEs underlying the energetics of the frontier orbitals. These features are then used to predict HOMO, LUMO, and E_g of “unknown” CPEs beyond the database. Finally, we discover a number of promising CPEs as hole transport materials in halide perovskite-based optoelectronic devices and as photocatalysts for water splitting.

RESULTS AND DISCUSSION

Construction of a CPE database

Multiple structural components of CPEs can be tuned independently and Fig. 1a illustrates such structural and chemical variability with CPE-K¹² as an example. The conjugated backbone of CPE-K consists of alternating electron-rich ($D\pi$) and electron-poor ($A\pi$) heterocycles; the intramolecular donor–acceptor interaction could yield lower bandgaps and charge-transfer states²⁵. The ionic functionality attached to the backbone comprises of an

¹Department of Physics and Astronomy, California State University Northridge, Northridge, California, USA. ²Department of Chemistry and Biochemistry, University of California, Santa Barbara, California, USA. ✉email: ganglu@csun.edu

Fig. 1 Molecular structure of CPE-K. Illustration of the molecular structures and building blocks of anionic CPEs, including the donor and acceptor units on the backbone, alkyl chains, anionic groups, and counter ions.

Fig. 2 Schematic diagram and results for HOMO, LUMO and E_g . **a** Illustration of LUMO/HOMO levels and E_g of a CPE; the HOMO/LUMO levels of its corresponding donor ($\text{HOMO}_D/\text{LUMO}_D$) and acceptor ($\text{HOMO}_A/\text{LUMO}_A$) units are also shown. **b** Summary of HOMO, LUMO levels, and E_g calculated with PBE functional as a function of the backbone index for 1296 anionic CPEs.

ionic group and an oppositely charged counter ion. The ionic group is responsible for CPE's solubility in polar solvents, including aqueous media. With a certain valence state, the size of the counter ion can also be adjusted²⁶. Finally, the distance between the ionic functionality and the electronically delocalized backbone can be modulated by the length of the alkyl side chain. The electrostatic interaction between the ionic functionality and charge carriers (polarons) on the conjugated backbone is known to play a crucial role in various applications of CPEs. Therefore, a complete characterization of CPEs requires the spanning of a 5-dimensional space of structural variations (donor, acceptor, alkyl chain length, ionic group, and counter ion) as illustrated in Fig. 1b.

The structural modularity of CPEs enables us to construct a database by expanding the five-dimensional space via high-throughput computation. At present, the database contains over 2000 CPEs whose properties (e.g., structural, electronic, optical, and dielectric properties) have been obtained via first-principles calculations (Supplementary Notes and Methods). Here, we focus on anionic CPEs with their anionic groups consisted of COO , SO_3 , and PO_3H , and the counter cations being Li, Na, K, and Cs. Three donor units (CPDT, P, PDT) and nine acceptor units (BT, PT, BBT, FBT, Ph, PhF, Py, PhCN, Tp) are combined to form 27 distinct

conjugated backbones for CPEs, as shown in Fig. 1b. The HOMO and LUMO levels of these donor and acceptor units are summarized in Supplementary Table 4. Four different lengths of the alkyl chain (C4, C6, C8, C10) are also included. In sum, 1296 ($=3 \times 4 \times 3 \times 9 \times 4$) anionic CPEs are considered in the present work.

Establishing structure–property relationships

HOMO/LUMO levels and E_g are three most important properties of CPEs in photovoltaic and photocatalytic applications, thus are the focus of our attention. HOMO and LUMO levels are often used as proxies for ionization potential and electron affinity, which can be measured experimentally. Note, that E_g is the HOMO-LUMO gap (or transport gap), shown schematically in Fig. 2a. We have systematically examined the HOMO/LUMO levels and E_g for all 1296 anionic CPEs and their values are summarized in Fig. 2b. The HOMO/LUMO levels are shifted relative to the vacuum level so that they can be compared among the CPEs. The vacuum level is defined as the potential energy at the vacuum of the simulation box (Supplementary Fig. 3). PBE functional was used in these calculations to extract the general trends. The horizontal axis

Fig. 3 Dependence of HOMO, LUMO and E_g on structural components. **a** Comparisons for HOMO/LUMO levels and E_g between CPEs* with C4 and C6 alkyl chains. **b** Charge density of HOMO and LUMO orbitals in CPE-K*. **c** HOMO level of CPE* as a function of HOMO level of the donor (HOMO_D). **d** LUMO level of CPE* as a function of LUMO level of the acceptor (LUMO_A). **e** E_g of CPE* as a function of LUMO_A-HOMO_D.

represents 27 different CPE backbones, each labeled by a backbone index (see Supplementary Table 5). For each backbone, there are 48 ($=3 \times 4 \times 4$) different sets of HOMO, LUMO, and E_g values, one for each distinct CPE of the same backbone but different combinations of the ionic group (3), counter ion (4) and length of alkyl chain (4). The first observation drawn from Fig. 2b is that for each backbone, the variation (~ 0.5 eV) in HOMO/LUMO level is much larger than that (~ 0.1 eV) in E_g . This suggests that in contrast to HOMO and LUMO, E_g is primarily determined by the backbone. However, the electrostatic force due to the ionic functionality can significantly modulate HOMO and LUMO for a given backbone. The second observation is that while many CPEs here have narrow bandgaps ($E_g < 2$ eV), some exhibit bandgaps wider than 2.5 eV. These wider gap CPEs all possess P as their donor unit (and Ph, PhF, and Py as their acceptor units) on the backbones. The lowest transport gap of these CPEs is 1.32 eV. Thus, in order to design NBCPEs, we need to discover donor/acceptor combinations in the backbone; tuning ionic functionality alone would not reach the goal. However, one can span a wide range of HOMO (-5.77 to -4.17 eV) and LUMO (-3.71 to -1.77 eV) levels by tuning the ionic functionality of the CPEs. Next, we will examine in detail how the various structural components can be tuned to modulate HOMO, LUMO, and E_g .

To examine the effect of the alkyl chain length on HOMO/LUMO and E_g , we consider a neutral CPE (termed CPE*), which is a CPE stripped of its ionic functionality. Two sets of CPEs* are studied here, each comprised of 27 CPEs*. One set has C4 chain length while the other has C6 chain length. We compare the values of HOMO/LUMO levels and E_g between the two sets, with each data point in Fig. 3a corresponding to a CPE* with the same backbone but different alkyl chain lengths (C6 vs. C4). Interestingly, there is a negligible energy difference between the two sets of CPEs*. The same results are also found between CPEs* with C8 and C10 alkyl chain lengths, suggesting that the chain length has a negligible effect on the energetics of the neutral CPEs.

We next examine the dependence of HOMO, LUMO, and E_g on the donor and acceptor units of the backbone. In Fig. 3b, the HOMO and LUMO charge densities of CPE-K* are displayed with

VESTA²⁷. It is observed that while the LUMO of CPE-K* resides primarily on the acceptor, the HOMO spreads on both the donor and the acceptor. Bader analysis was used to compute the charge distribution on the donor and acceptor units by removing or adding an electron to CPE-K*. With the removal of the electron, the positive charges on the donor and the acceptor are 0.63e and 0.37e, respectively, and when adding an electron, the negative charges on the donor and the acceptor are -0.31 e and -0.69 e, respectively. Thus, it is reasonable (but not precise) to correlate the HOMO level of CPE* to the HOMO level of the donor (HOMO_D) and the LUMO level of CPE* to the LUMO level of the acceptor (LUMO_A). In Fig. 3c, we plot the HOMO level of CPE* vs. HOMO_D for the three donors CPDT, P, PDT, and approximate linear relationships are found for all nine acceptor units. Thus, there exist a strong positive correlation between the HOMO level of a neutral CPE and the HOMO level of its donor unit. Similarly, as shown in Fig. 3d for each donor, there is a positive correlation between LUMO of CPE* and LUMO_A. As $E_g = \text{LUMO} - \text{HOMO}$, a positive but weaker correlation is also identified between E_g and LUMO_A-HOMO_D in Fig. 3e. While not further pursued in this paper, these correlations can provide valuable guidance in designing donor/acceptor combinations for targeted applications. For example, donors of low HOMO levels combined with acceptors of low LUMO levels would likely to form CPEs with low bandgaps.

Next, we explore how the electrostatic force may affect the HOMO/LUMO levels of CPEs. To this end, we focus on CPEs with a fixed backbone (CPDT-BT) and anionic group (COO), but with different alkyl chain lengths (C4, C6, C8, C10) and counter ions (Li, Na, K, Cs). The varied electronegativities of the counter ions and Coulomb interaction distances due to the chain lengths modulate the electrostatic force that the polaron in the conjugated backbone may experience. In Fig. 4a, we display the variation of the HOMO level as a function of the alkyl chain length and the counter ion. The dashed horizontal line represents the HOMO level of the neutral CPE* with the same CPDT-BT backbone and C4 alkyl chain. Three general trends are observed: (1) for a given chain length, the HOMO level decreases monotonically with the

Fig. 4 Dependence of HOMO on alkyl chain length and counter ion. **a** The HOMO level of CPEs (with CPDT-BT backbone and COO ionic group) as a function of alkyl chain length (C4, C6, C8, C10) and counter ion (Li, Na, K Cs). The dashed horizontal line is the HOMO level of corresponding CPE* (CPDT-BT). **b** Schematic picture illustrating the polarization of a π -orbital on the backbone due to the presence of anionic group or a counter ion²⁸. The yellow lobes indicate the π -electron distribution and the dashed contour represents the original π -orbital before the polarization. The positive and negative ions are shown by + and – signs.

electronegativity of the counter ion; the smaller the electronegativity, the higher the HOMO. (2) For a given counter ion or electronegativity, the HOMO level decreases monotonically with the chain length; the longer the alkyl chain, the lower the HOMO level. (3) The HOMO level of the neutral CPE* (HOMO*) represents an asymptotic limit of the HOMO level for all anionic CPEs; in the limit of an infinite alkyl chain length, $\text{HOMO} = \text{HOMO}^*$. In this case, the electrostatic force vanishes.

All these trends can be rationalized from an electrostatic point of view. The electrostatic interaction between a charged ion (or ionic group) and a π -orbital on the backbone is illustrated in Fig. 4b. The π -orbital of the neutral CPE* is represented by the dashed contour. When a CPE is surrounded by charged ions, its π -electrons on the backbone are attracted by the positive ions and lower their energies, while the negative ions repel the π -electrons and increase their energies²⁸. This effect is purely electrostatic and becomes more pronounced as the ion-backbone distance decreases. While both the anionic group (COO^-) and the counter ion (Li, Na, K, Cs) could polarize the π -orbital, the anionic group tends to have a stronger effect than the counter ion since it is closer to the backbone. The smaller the electronegativity of the counter ion, the more the charge transfer from the counter ion to the anionic group; thus, the anionic group is more negatively charged, leading to a higher HOMO level. In the same vein, the shorter the chain length, the stronger the polarization, and the higher the HOMO level. The polarization disappears in a neutral CPE, thus $\text{HOMO}^* < \text{HOMO}$.

The same general trends also hold in CPEs with other backbones and ionic groups, as shown in Supplementary Fig. 7. But there are also notable exceptions. For example, with the smallest counter ion (Li) and SO_3 group, the HOMO level is actually lower than HOMO^* . Additionally, CPEs with the shortest chain length (C4) and smaller counter ions (Li and Na) exhibit lower HOMO level as shown in Supplementary Fig. 8. These anomalies arise because the counter ions are situated closer to the backbone than the ionic group. Since the smaller counter ions (Li and Na) are more mobile, it is easier for them to circumvent the anionic group and to come near the backbone. This scenario is more likely to occur when the chain length is short as the counter ion can bond with two adjacent anionic groups simultaneously, facilitating the counter ion to draw closer to the backbone. The LUMO level exhibits similar trends as the HOMO (Supplementary Fig. 9) which is expected since E_g shows weak dependence on the electrostatic force.

The preceding analysis has demonstrated that the frontier orbitals and E_g of CPEs can be modulated by the HOMO/LUMO levels of the donor and acceptor units in the backbone, the alkyl chain length, and the electronegativity of the ionic functionality. However, owing to complex correlations among these quantities,

concrete structure–property relationships, necessary for the rational design of CPEs, are difficult to establish. Thus, in the following, we resort to machine learning to derive appropriate structure–property relationships for predicting CPE properties. Since each ionic group has an internal structure, the electronegativity alone would not be sufficient to represent the characteristics of the ionic group. Hence, we also include its coordination number (CN) in the machine learning modeling. Specifically, the CN of each of the three ionic groups (COO , SO_3 and PO_3H) refers to the number of bonds on the C atom in COO (CN = 2, with bonds to 2 oxygen atoms), S atom in SO_3 (CN = 3, with bonds to 3 oxygen atoms) and P atom in PO_3H (CN = 3, with bonds to 2 oxygen atoms and 1 OH), respectively. We have thus incorporated eight features, including HOMO and LUMO of the donor and acceptor units (HOMO_D , LUMO_D , HOMO_A , LUMO_A), the alkyl chain length (L), the electronegativity of the anionic group (χ^-) and the counter cation (χ^+), and the coordination number of the anionic group (CN) in the machine learning analysis, and their values are summarized in Tables S4 and S6.

The Pearson correlation coefficient (p) is commonly employed to represent a linear relationship between a given feature and a property in machine learning analysis, and it is defined in the following:²⁹

$$p = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (1)$$

where $[X_i]$ and $[Y_i]$ are corresponding datasets for the feature and property, respectively, and n is the number of CPEs in the dataset. $\bar{X}(\bar{Y})$ is the averaged value of the feature (property). A greater positive (or negative) p value indicates a stronger correlation (or anti-correlation) between the property Y and the feature X . While the Pearson correlation coefficient analysis is simple and transparent, it is not always accurate specially when nonlinear relationships between features and properties are important. To remedy this deficiency, we also perform Random Forest Regression (RFR) in our analysis to complement the Pearson coefficient. RFR is a supervised learning algorithm that utilizes ensemble learning method for regression, and thus is more reliable than the linear analysis. With RFR, we can rank the relative importance of each feature³⁰, represented by the feature importance variable Q (see Supplementary Methods). In Fig. 5, we summarize the Pearson correlation coefficients (p) and feature importance (Q) for the eight features on HOMO/LUMO levels and E_g . On the one hand, the results in Fig. 5a confirm what we have already known, i.e., E_g depends primarily on the backbone, with HOMO_D and LUMO_A playing the most important roles. On the other hand, the machine learning analysis can also reveal more concrete and sometimes hidden correlations. For instance, LUMO_D and to a less

Fig. 5 Feature analysis via machine learning. Correlation coefficients (p) and feature importance (Q) as a function of the structural features for E_g (a), HOMO (b) and LUMO (c) levels.

extent HOMO_A are found to be important to E_g . While LUMO_D , HOMO_A and LUMO_A are correlated to E_g , HOMO_D is anti-correlated to E_g . As discussed earlier, HOMO of a CPE is expected to depend on HOMO_D . Interestingly, however, we also uncover that the HOMO of a CPE depends on strongly on LUMO_D and HOMO_A , as shown in Fig. 5b. In fact, from the RFR analysis, HOMO_A is revealed as important as HOMO_D in modulating the HOMO of the CPE. This revelation is actually consistent with Fig. 3b which shows that the HOMO of CPE-K* spreads on both the donor and acceptor units. In addition, as displayed in Fig. 5c, the linear Pearson analysis grossly overestimates the importance of HOMO_A in determining the LUMO of a CPE, as compared to the RFR analysis. To our surprise, the donor unit appears to have negligible influence on the LUMO level of the CPE. The RFR analysis indicates that the LUMO level depends almost exclusively on LUMO_A while the HOMO has roughly an equal dependence (but with different signs) on HOMO_D , LUMO_D and HOMO_A . The latter reflects the significant interaction/charge transfer between the donor and the acceptor on the backbone. Although the ionic group, chain length and counter ion individually plays a minor role in determining the HOMO/LUMO levels, collectively they can change the HOMO/LUMO levels as much as 0.5 eV, thus cannot be discounted in the machine learning analysis. Among them, the alkyl chain length (L) and the electronegativity of the ionic group (χ^-) are less important to the HOMO/LUMO levels. In particular, χ^- has negligible influence on them, thus is removed from the further analysis. In contrast, the coordination number (CN) of the ionic group outranks L , χ^- and χ^+ in importance, suggesting that the design of the ionic group should not be taken lightly.

Predicting CPE properties via machine learning

Our goal in machine learning is to provide guidance for materials design, i.e., to discover CPEs for target applications. To accomplish this goal, we need to uncover material features that can be directly linked to the molecular structures of CPEs. In addition, we aim to predict materials properties (HOMO, LUMO and E_g) for a given CPE structure. Progress is made on both tasks here. Since the HOMO/LUMO of the donor and acceptor units are not directly linked to the molecular structures of CPEs, they are removed from the feature list. For the first task, we include the following revised features to describe the donor and acceptor units (Supplementary Table 6) in the machine learning analysis: the degree of unsaturation (defined as the number of rings and π bonds) in the donor and acceptor units (DU_D , DU_A), the differential electronegativity between a non-carbon atom and a carbon atom on the ring of the donor and acceptor unit (ΔD_1 , ΔA_1) (i.e., ΔD_1 and ΔA_1 are zero if all atoms on the ring are carbon), and the differential electronegativity between a substituent group and a hydrogen atom in the donor and acceptor unit (ΔD_2 , ΔA_2) (i.e., ΔD_2 and ΔA_2 are zero if there is no substituent group). Since ΔD_2 is the

same for the three donors in the database, this feature is removed in the following analysis, but should be included in general. In addition to the five features, the same ones used to represent the ionic functionality (CN, L , χ^- and χ^+) are included in the analysis. More information on the degree of unsaturation and differential electronegativity can be found in Supplementary Methods.

In Fig. 6a–c, we present the Pearson correlation coefficient p and the feature importance Q for E_g and HOMO/LUMO levels using the revised features. As expected, the five features characterizing the donor and acceptor units are all important to E_g . While DU_D is positively correlated to E_g , other features (DU_A , ΔD_1 , ΔA_1 and ΔA_2) are anti-correlated to E_g . Thus, in order to lower E_g , one needs to reduce the degree of unsaturation on the donor. As shown in Supplementary Table 4, with the degree of unsaturation (DU_D) decreasing from P, to PDT and to CPDT, the corresponding HOMO_D increases. According to Fig. 5b, c, the HOMO of CPEs is positively correlated to HOMO_D while the LUMO of CPEs has negligible dependence on HOMO_D . Thus, the decrease of DU_D would result in a reduction of the bandgap. Other ways can be used to reduce E_g are increasing the degree of unsaturation on the acceptor, increasing the electronegativity of the non-carbon atoms on the ring, and increasing the electronegativity of the substituent group. Three most important features for HOMO are DU_D , ΔD_1 and ΔA_2 . Thus, one can increase the HOMO level by lowering the degree of unsaturation on the donor and the electronegativity of the substituent and by increasing the electronegativity of non-carbon atoms on the ring. The most important features to the LUMO level are all related to the acceptor; and to increase LUMO, one can reduce the degree of unsaturation, the electronegativity of non-carbon atoms on the ring and the substituent group of the acceptor. These structure–property relations can provide valuable guidance for rational design of CPEs.

For the second task, we have used machine learning to predict HOMO/LUMO levels for CPEs beyond the database. Here, the Support Vector Regression (SVR) method as implemented in “scikit-learn”³¹ was used for the regression of the HOMO and LUMO levels (see Supplementary Methods). Specifically, a 10-fold cross-validated SVR approach with the grid search was used and its performance was evaluated by the mean-absolute-percent-error (MAPE), the root-mean-square-error (RMSE) and the correlation coefficient (R^2) for the prediction of HOMO/LUMO levels, as shown in Supplementary Fig. 10. We find that the performance of the SVR regression model is stable and robust against different randomizations. In Fig. 6d, e, we display the comparisons of HOMO/LUMO levels between the machine learning predictions and the first-principles results, with 1046 CPEs in the training set and 250 CPEs in the test set. It is found that the machine learning model performs well with MAPE less than 2.1%, RMSE less than 0.07 eV and R^2 larger than 0.97 (Table 1). To assess the ability of

Fig. 6 Prediction of HOMO and LUMO via machine learning. Correlation coefficients (p) and feature importance (Q) as a function of the revised structural features for E_g (a), HOMO (b) and LUMO (c) levels. The comparisons between the DFT and the machine learning results for HOMO (d) and LUMO (e) levels, respectively.

Table 1. Summary of MAPE (%), RMSE (eV) and R^2 for the training, test and verification datasets using the revised features.

	MAPE			RMSE			R^2		
	training	test	verification	training	test	verification	training	test	verification
HOMO	0.92	0.97	1.16	0.05	0.06	0.09	0.97	0.97	0.91
LUMO	2.07	2.07	3.63	0.06	0.07	0.14	0.97	0.98	0.90

the machine learning model in predicting HOMO/LUMO levels beyond the database, we also constructed a verification dataset (Supplementary Tables 6 and 7, Supplementary Fig. 11), which has not been previously used in the training and test sets. With all known CPEs (1296) in the database as the training set, we can predict HOMO/LUMO levels of CPEs in the verification set and the results are summarized in Fig. 6d, e and Table 1. In general, we find that the machine learning model can provide reasonable estimate for the HOMO/LUMO levels, and the revised structural features can be used as descriptors in predicting the properties of unknown CPEs.

Screening CPEs for optoelectronic and photocatalytic applications

CPEs have found a wide range of applications from optoelectronics to photocatalysts. However, rational design of CPEs has yet to benefit from data-centric research. In this work, we use machine learning to screen/discover CPEs for two selective applications. For the first application, we focus on the design of hole transport materials (HTMs) in metal-halide perovskite based optoelectronic devices, such as photovoltaics and light-emitting diodes^{5,7,10,32–34}. For the second application, we aim to design photocatalysts for overall water splitting^{13,14,35}. Both applications require the design of CPEs with appropriate HOMO, LUMO, and E_g values, along with other relevant properties.

For the first application, the ideal HTMs should have appropriate energy levels so that the holes can be transferred from the active layer to the electrode efficiently while the electrons are blocked^{32,36}. The solution-processed pH-neutral CPEs have excellent compatibility and wetting property, which enables the formation of a stable, defect-suppressed and uniform interface with the perovskite active layer³⁷. Here, three design criteria are considered for the screening of CPEs: stability, energy level alignment and hole mobility. Specifically, the CPEs should have a higher energetic stability than CPE-K, which has been used as HTMs in conjunction with perovskites. We require the candidate to have an atomization energy lower than that (-0.239 eV) of CPE-K. Note that this is not a critical criterion and helps to reduce the number of candidates that we have to consider. Secondly, we require that the HOMO level of the candidate be slightly higher than the valence band maximum (VBM) of the perovskite, but lower than the VBM of the ITO (indium tin oxide) electrode (Fig. 7). Thus, photogenerated holes in the active layer can be efficiently extracted. At the same time, in order to block the electron transfer from the perovskite to the HTM, the LUMO of the candidate should be lower than that of the perovskite (~ 3.7 eV). Thirdly, we require that the candidate should have a lower reorganization energy than CPE-K (0.186 eV). This requirement derives from the consideration of charge-transfer rates (K) in HTMs, which can be

estimated based on the Marcus theory:^{38,39}

$$K = \left(\frac{V^2}{\hbar} \right) \left(\frac{\pi}{\lambda k_B T} \right)^2 \exp(-\lambda/4k_B T) \quad (2)$$

Here $\hbar$, T and k_B are reduced Planck constant, temperature and Boltzmann constant, respectively. λ is the reorganization energy and V is the electronic coupling. For polymers including CPEs, the reorganization energy is often the dominant factor on the charge-transfer rate, and the smaller the reorganization energy, the higher the charge-transfer rate⁴⁰. The reorganization energy for the holes in CPEs can be estimated as follows:⁴⁰

$$\lambda = (E_0^+ - E_+^+) + (E_+^0 - E_0^0) \quad (3)$$

where E_0^+ is the energy of the cation in the optimized neutral state structure; E_+^+ is the energy of the cation in the optimized cationic structure; E_+^0 is the energy of the neutral molecule in the optimized cationic structure, and E_0^0 is the energy of the neutral molecule in the optimized neutral state structure. Thus, the candidate should have λ less than that (0.186 eV) of CPE-K. Finally, the candidate should have polar functionality, which is a condition satisfied by all CPEs in the database. Among ~1300 anionic CPEs in the database, 72 candidates met the above criteria. Among them, 5 candidates with the lowest reorganization energies are included in Table 2. It is interesting to note that the CPE with a low reorganization energy generally has a longer alkyl chain length than CPE-K (C4). As shown in Supplementary Fig. 12, for a given backbone (CPDT-BT) and functionality (SO₃K), as the alkyl chain length increases, the hole reorganization energy decreases.

Fig. 7 Energy diagrams for HTMs and photocatalysts. Energy diagrams for the hole transport materials (HTMs, left) and the photocatalysts for water splitting (right). The energy windows for the conduction band (up) and the valence band (bottom) of CPEs are shown in green. The corresponding energy windows for ITO and the perovskite active layer are shown in blue and brown, respectively. The dash lines indicate the redox potentials of water.

Therefore, we propose to increase the alkyl chain length on the basis of CPE-K for synthesizing more efficient HTMs.

CPEs have also received recent attention as photocatalysts for water splitting. Some of the advantages that CPEs offer as photocatalysts include turnability, low-cost with earth abundant and non-toxic materials, and water solubility^{13,14,35,41}. In addition, the polar side chains of CPEs can interact strongly with metal electrodes and/or metal co-catalysts, yielding efficient charge extraction/transfer to the electrodes and the co-catalysts⁴². However, insofar the photocatalytic activities of CPEs remain low.

To serve as photocatalysts for overall water splitting, the HOMO and LUMO levels of CPEs should straddle across the redox potentials of water⁴³. As the overpotential for an efficient hydrogen evolution reaction (HER) and oxygen evolution reaction (OER) catalyst is about 0.2 eV^{44,45}, we require that the candidate to have its HOMO level lower than -5.87 eV and LUMO level higher than -4.24 eV, respectively to ensure that the photo-induced hole and electron can overcome the overpotential⁴⁶. In addition, the bandgap of the candidate should be less than 2.4 eV for efficient light absorption. Finally, we require the candidate to have an atomization energy lower than that of CPE-K. Based on these criteria, 17 candidates are identified for photocatalytic water splitting, and 5 of them with the lowest E_g are included in Table 2. We find that all promising candidates have "P" molecule as the donor unit. This observation is consistent with the fact that all CPE photocatalysts for water splitting reported so far contain "P" in their backbones. In addition, compared to the CPEs that can catalyze only HER (half-water splitting)³⁵, our predicted candidates have deeper HOMO levels in the acceptors (e.g., FBT and PhCN), which enable overall water splitting and lower bandgaps. Thus, to improve the performance of CPEs as water splitting photocatalysts, one should focus on engineering acceptor units, such as by tuning their ΔA_1 and ΔA_2 .

Although the machine learning and data driven approach presented here are powerful tools in discovering CPEs for target applications, one has to bear in mind their limitations too. More data (in particular, experimental data) are critically needed to improve the fidelity of the predictions. More sophisticated theoretical models have to be developed to capture crucial physical/chemical processes, including charge transport, the role of solvent and defects, interfacial wettability, lifetime of charge carriers, etc.

In conclusion, we have constructed a first-principles database on CPEs, which allows us to systematically examine how HOMO/LUMO levels and E_g depend on the structural components of CPEs, including the donor/acceptor units on the backbone, the alkyl chain length, the ionic group and the counter ion. General trends correlating HOMO/LUMO and E_g with the structural components are established and can be rationalized in terms of the electrostatic interaction between the ionic functionalities and polarons on the CPE backbone. By means of machine learning, we have uncovered material features important to frontier orbitals and E_g and the structure–property relations for rational design of CPEs. These features are then used as descriptors in the machine learning to predict HOMO/LUMO levels and E_g of "unknown" CPEs

Table 2. Selected candidates for HTMs and water splitting photocatalysts. The values of HOMO, LUMO, reorganization energy (λ) and E_g are displayed (all in eV).

HTMs				Photocatalysts			
CPE	HOMO	LUMO	λ	CPE	HOMO	LUMO	E_g
CPDT-PhF-C8-SO ₃ Na	-5.136	-2.845	0.131	P-FBT-C4-SO ₃ Li	-5.910	-3.881	2.029
PDT-Ph-C6-SO ₃ Na	-5.185	-2.654	0.132	P-FBT-C6-SO ₃ Li	-5.877	-3.841	2.036
CPDT-PhF-C10-SO ₃ Na	-5.145	-2.851	0.134	P-PhCN-C4-PO ₃ HNa	-6.034	-3.781	2.253
CPDT-FBT-C8-SO ₃ K	-5.184	-3.629	0.138	P-PhCN-C4-PO ₃ HK	-6.016	-3.732	2.284
CPDT-PhF-C10-SO ₃ K	-5.103	-2.808	0.140	P-PhCN-C4-PO ₃ HCS	-6.012	-3.727	2.285

beyond the database. Finally, we propose relevant design criteria based on which promising CPEs are identified as the hole transport materials in halide perovskite optoelectronic devices and as photocatalysts for water splitting. Our work represents one of the data-centric research dedicated to CPEs and could accelerate the discovery and development of CPEs for optoelectronic and photocatalytic applications.

METHODS

Computational models and details

Our first-principles calculations are based on density functional theory (DFT) with planewave bases as implemented in Vienna Ab initio Simulation Package (VASP)^{47–50}. The projector-augmented-wave method was used to describe the pseudopotentials. Both Perdew–Burke–Ernzerhof (PBE) and HSE06 hybrid functionals^{51–53} were considered, with the former used on the entire CPE database and the latter on a subset of the database (Supplementary Table 1). As conjugated polymers, CPEs could adopt various conjugation lengths. Obviously, we cannot include all possible conjugation lengths for all CPEs (~1300) in the database. Instead, we focus on CPE oligomers with a single repeating conjugation unit. However, we have also considered a subset of CPEs with infinite conjugation lengths. In particular, we have examined the variation of HOMO/LUMO/ E_g energies as a function of the conjugation length for CPE-K (Supplementary Fig. 4). To model CPEs with a single unit of conjugation length, we included a vacuum of 15 Å in each direction of the computational cells and found it sufficient to isolate periodic images of CPE molecules. Due to the steric hindrance, a fully extended and trans-conformational alkyl chain is energetically more stable than other conformations, thus is adopted as the initial structure for each CPE molecule⁵⁴. We then relax the initial molecular structure to obtain its equilibrium conformation at 0 K. For a few select CPEs, we have also performed ab initio molecular dynamics at 300 K to examine whether the relaxed equilibrium structures remain stable (i.e., whether there are drastic conformational changes to the equilibrium structures). All equilibrium structures examined are found to be thermodynamically stable and can be used as bases to understand the structure–property relationships. Although local minima exist in the molecular structures of CPEs, selective analysis finds them to have minor influence on the results. An energy cutoff was set as 500 eV and the convergence threshold for the energy and force were 10^{-5} eV and 0.02 eV/Å, respectively. The inclusion of van der Waals dispersion correction at the level of PBE+D3⁵⁵ is found to have negligible effect on the energetics and the structures of CPEs (Supplementary Fig. 5, Supplementary Table 2).

It is known that PBE functional generally underestimates E_g compared to HSE06 functional and experimental values. Interestingly, as shown in Supplementary Table 1, the ratios between the corresponding HSE06 and PBE values for CPEs are more or less constant; in particular, the ratio for E_g is ~1.4 and for HOMO level is ~1.1. Since the HSE06 calculation is much more expensive than PBE, one could save tremendous computational time by scaling the PBE values (of E_g and HOMO) by the constant ratios to approximate the corresponding HSE06 values without actually performing the HSE06 calculations on the entire set of CPEs in the database. On the other hand, it is also known and confirmed by our calculations that E_g for CPEs with infinite conjugation lengths is always lower than that for CPEs with a single conjugation unit as well as the experimental value. Actually, modeling CPEs with a single conjugation unit by PBE functional turns out to yield good agreements to the experimental values on E_g as shown in Supplementary Table 3. As for the HOMO level, we find that HSE06 functional with a single unit of conjugation length yields results consistent with the experiments. Therefore, when screening CPEs for target applications, PBE functional will be used for E_g while HSE06 functional will be used for HOMO level, both with a single unit of conjugation length.

DATA AVAILABILITY

The data that support this work are available in the article and Supplementary Information file. Further raw data can be found at the CPE database: <https://cpegenome.com> and are available from the corresponding author (G.L.) upon request.

CODE AVAILABILITY

Scripts for generating CPE molecular structures are available at: <https://cpegenome.com>.

Received: 8 January 2021; Accepted: 19 April 2021;

Published online: 20 May 2021

REFERENCES

- Patil, A. O., Ikenoue, Y., Wudl, F. & Heeger, A. J. Water soluble conducting polymers. *J. Am. Chem. Soc.* **109**, 1858–1859 (1987).
- Hoven, C. V., Garcia, A., Bazan, G. C. & Nguyen, T. Q. Recent applications of conjugated polyelectrolytes in optoelectronic devices. *Adv. Mater.* **20**, 3793–3810 (2008).
- Lee, W., Seo, J. H. & Woo, H. Y. Conjugated polyelectrolytes: a new class of semiconducting material for organic electronic devices. *Polymer* **54**, 5104–5121 (2013).
- Jiang, H., Taranekekar, P., Reynolds, J. R. & Schanze, K. S. Conjugated polyelectrolytes: synthesis, photophysics, and applications. *Angew. Chem. Int. Ed.* **48**, 4300–4316 (2009).
- Huang, F., Wu, H. & Cao, Y. Water/alcohol soluble conjugated polymers as highly efficient electron transporting/injection layer in optoelectronic devices. *Chem. Soc. Rev.* **39**, 2500–2521 (2010).
- Zhou, H. et al. Solution-processed pH-neutral conjugated polyelectrolyte improves interfacial contact in organic solar cells. *ACS Nano* **9**, 371–377 (2015).
- Choi, H. et al. Conjugated polyelectrolyte hole transport layer for inverted-type perovskite solar cells. *Nat. Commun.* **6**, 7348 (2015).
- Han, Y. W. et al. Vertical phase separation for highly efficient organic solar cells incorporating conjugated-polyelectrolytes. *Adv. Mater. Interfaces* **6**, 1801396 (2019).
- Seo, J. H., Namdas, E. B., Gutacker, A., Heeger, A. J. & Bazan, G. C. Solution-processed organic light-emitting transistors incorporating conjugated polyelectrolytes. *Adv. Funct. Mater.* **21**, 3667–3672 (2011).
- Lee, B. R. et al. Conjugated polyelectrolytes as efficient hole transport layers in perovskite light-emitting diodes. *ACS Nano* **12**, 5826–5833 (2018).
- Hu, Z., Zhang, K., Huang, F. & Cao, Y. Water/alcohol soluble conjugated polymers for the interface engineering of highly efficient polymer light-emitting diodes and polymer solar cells. *Chem. Commun.* **51**, 5572–5585 (2015).
- Mai, C. K. et al. Side-chain effects on the conductivity, morphology, and thermoelectric properties of self-doped narrow-band-gap conjugated polyelectrolytes. *J. Am. Chem. Soc.* **136**, 13478–13481 (2014).
- Wu, Y. et al. Substituent regulation improves photocatalytic hydrogen evolution of conjugated polyelectrolytes. *ACS Mater. Lett.* **1**, 620–627 (2019).
- Yang, X. et al. Water-soluble conjugated molecule for solar-driven hydrogen evolution from salt water. *Adv. Funct. Mater.* **29**, 1808156 (2019).
- Cui, Q. & Bazan, G. C. Narrow band gap conjugated polyelectrolytes. *Acc. Chem. Res.* **51**, 202–211 (2018).
- Kirchhofer, N. D., McCuskey, S. R., Mai, C.-K. & Bazan, G. C. Anaerobic respiration on self-doped conjugated polyelectrolytes: impact of chemical structure. *Angew. Chem. Int. Ed.* **129**, 6619–6622 (2017).
- Mai, C. K. et al. Facile doping of anionic narrow-band-gap conjugated polyelectrolytes during dialysis. *Angew. Chem. Int. Ed.* **52**, 12874–12878 (2013).
- Mai, C. K. et al. Electrical properties of doped conjugated polyelectrolytes with modulated density of the ionic functionalities. *Chem. Commun.* **51**, 17607–17610 (2015).
- Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 1–7 (2016).
- Oses, C., Toher, C. & Curtarolo, S. Data-driven design of inorganic materials with the automatic flow framework for materials discovery. *MRS Bull.* **43**, 670–675 (2018).
- Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
- Sendek, A. D. et al. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem. Mater.* **31**, 342–352 (2018).
- Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).
- Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Mullen, K. & Pisula, W. Donor-acceptor polymers. *J. Am. Chem. Soc.* **137**, 9503–9505 (2015).

26. Cui, Y. et al. The critical role of anode work function in non-fullerene organic solar cells unveiled by counterion-size-controlled self-doping conjugated polymers. *Chem. Mater.* **30**, 1078–1084 (2018).
27. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).
28. Li, Z., Zhang, X., Lu, G. & Nguyen, T.-Q. First-principles study of electron mobility in cationic and anionic conjugated polyelectrolytes. *J. Phys. Chem. C* **116**, 1205–1210 (2011).
29. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
30. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recogn. Lett.* **31**, 2225–2236 (2010).
31. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
32. Lee, J. et al. Achieving large-area planar perovskite solar cells by introducing an interfacial compatibilizer. *Adv. Mater.* **29**, 1606363 (2017).
33. Jung, E. D. et al. Multiply charged conjugated polyelectrolytes as a multi-functional interlayer for efficient and scalable perovskite solar cells. *Adv. Mater.* **32**, 10 (2020).
34. Rakstys, K., Igci, C. & Nazeeruddin, M. K. Efficiency vs. stability: dopant-free hole transporting materials towards stabilized perovskite solar cells. *Chem. Sci.* **10**, 6748–6769 (2019).
35. Hu, Z. et al. Highly efficient photocatalytic hydrogen evolution from water-soluble conjugated polyelectrolytes. *Nano Energy* **60**, 775–783 (2019).
36. Urieta-Mora, J., Garcia-Benito, I., Molina-Ontoria, A. & Martin, N. Hole transporting materials for perovskite solar cells: a chemical approach. *Chem. Soc. Rev.* **47**, 8541–8571 (2018).
37. Lee, S. et al. Conjugated polyelectrolytes as multifunctional passivating and hole-transporting layers for efficient perovskite light-emitting diodes. *Adv. Mater.* **31**, e1900067 (2019).
38. Marcus, R. A. Electron transfer reactions in chemistry. Theory and experiment. *Rev. Mod. Phys.* **65**, 599 (1993).
39. Marcus, R. A. Chemical and electrochemical electron-transfer theory. *Annu. Rev. Phys. Chem.* **15**, 155–196 (1964).
40. Brédas, J.-L., Beljonne, D., Coropceanu, V. & Cornil, J. Charge-transfer and energy-transfer processes in π -conjugated oligomers and polymers: a molecular picture. *Chem. Rev.* **104**, 4971–5004 (2004).
41. Bai, Y. Q., Hu, Z. C., Jiang, J. X. & Huang, F. Hydrophilic conjugated materials for photocatalytic hydrogen evolution. *Chem. -Asian J.* **15**, 1780–1790 (2020).
42. Leem, G. et al. Photocathode chromophore–catalyst assembly via layer-by-layer deposition of a low band-gap isoindigo conjugated polyelectrolyte. *ACS Appl. Energy Mater.* **1**, 62–67 (2018).
43. Li, Y., Li, Y.-L., Sa, B. & Ahuja, R. Review of two-dimensional materials for photocatalytic water splitting from a theoretical perspective. *Catal. Sci. Technol.* **7**, 545–559 (2017).
44. Qiu, Z., Tai, C.-W., Niklasson, G. A. & Edvinsson, T. Direct observation of active catalyst surface phases and the effect of dynamic self-optimization in NiFe-layered double hydroxides for alkaline water splitting. *Energy Environ. Sci.* **12**, 572–581 (2019).
45. Suryanto, B. H. R., Wang, Y., Hocking, R. K., Adamson, W. & Zhao, C. Overall electrochemical splitting of water at the heterogeneous interface of nickel and iron oxide. *Nat. Commun.* **10**, 5599 (2019).
46. Valdes, A. & Kroes, G.-J. First principles study of the photo-oxidation of water on tungsten trioxide (WO₃). *J. Chem. Phys.* **130**, 114701 (2009).
47. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
48. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comp. Mater. Sci.* **6**, 15–50 (1996).
49. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
50. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
51. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
52. Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
53. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
54. Strobl, G. (ed.) *The Physics of Polymers* Ch. 2, 13–15 (Springer, 2007).
55. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A. consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **132**, 154104 (2010).

ACKNOWLEDGEMENTS

The work was supported by the U.S. National Science Foundation (DMREF-1922042).

AUTHOR CONTRIBUTIONS

G.L. designed the project. Y.W. carried out the computation and data analysis with the assistance of X.Z., G.L. supervised the research and wrote the manuscript with Y.W. All authors discussed the results and were involved in the writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00541-5>.

Correspondence and requests for materials should be addressed to G.L.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021