Harnessing DFT and Machine Learning for Accurate Optical Gap Prediction in Conjugated Polymers

Bin Liu[1,2†], Yunrui Yan[1,2†], and Mingjie Liu[1,2*]

[1]Department of Chemistry, University of Florida, Gainesville, FL 32611, United States.

[2]Quantum Project Theory, University of Florida, Gainesville, FL 32611, United States.

[†]These authors contributed equally to this work.

[*]Corresponding author. E-mail: mingjieliu@ufl.edu

August 13, 2024

Abstract

Conjugated polymers (CPs), characterized by alternating σ and π bonds, have attracted significant attention for their diverse structures and adjustable electronic properties. However, predicting the optical band gap (E_{gap}^{exp}) of CPs remains challenging. This study presents a rational model that integrates density functional theory (DFT) calculation with a data-driven machine learning (ML) approach to predict the experimentally measured E_{gap}^{exp} of CPs, using 1096 data points. Through alkyl side chain truncation and conjugated backbone extension, the modified oligomers effectively capture the electronic properties of CPs, significantly improving the correlation between the DFT-calculated HOMO-LUMO gap ($E_{gap}^{oligomer}$) and E_{gap}^{exp} (R^2=0.51) compared to the unmodified side-chain-containing monomers (R^2=0.15). Moreover, we trained six ML models with two categories of features as input: $E_{gap}^{oligomer}$ to represent the extended backbone and molecular features of unmodified monomers to capture the alkyl-side-chain effect. The best model, XGBoost-2, achieved an R^2 of 0.77 and an MAE of 0.065 eV for predicting E_{gap}^{exp}, falling within the experimental error margin of ∼0.1 eV. We further validated XGBoost-2 on a dataset of 227 newly synthesized CPs collected from literature without further retraining. Notably, XGBoost-2 exhibits both excellent interpolation for BT-, BTA-, QA-, DPP-, and TPD-based CPs, and exceptional extrapolation for PDI-, NDI-, DTBT-, BBX-, and Y6-based CPs, which are attributed to the integration of DFT methods with rationally designed oligomer structures. For the first time, we demonstrated a novel and effective strategy combining quantum chemistry calculations with ML modeling for accurate and efficient prediction of experimentally measured fundamental properties of CPs. Our study paves the way for the accelerated design and development of high-performance CPs in photoelectronic applications.

Keywords: Conjugated polymers, Experimental optical gap, Density functional theory, Machine learning, Photoelectronics

1 Introduction

Conjugated polymers (CPs) are organic macromolecules composed of electron donor and electron acceptor units linked by carbon-carbon bonds. The alternation of σ and π bonds along the backbone chain of CPs enables the delocalization of π-electrons, forming a semiconductor band structure and thus endowing CPs with exceptional optical and electronic properties.[1–3] These characteristics can be effectively tuned through a variety of material engineering strategies, such as the combination of various electron donator and electron acceptor aromatic units,[4, 5] halogenation,[6] the introduction of non-covalent intra- and inter-molecular interactions,[7, 8] and modifications to alkyl side chains.[9] Owing to their structural diversity, facile synthesis, ease of chemical modification and functionalization, excellent photo-physical properties, and relatively low cost, CPs have been extensively explored for a wide range of applications in optoelectronic devices, electrochemical sensors and transistors, drug delivery systems and bio-medical applications.[10–13] So far, hundreds of thousands of CPs are available, but the scientific community predominantly relies on a laborious trial-and-error approach for the discovery, design, and optimization of CP materials, resulting in a substantial number of unexplored structures. The relationships between the structures of CP materials and their electronic properties are complex and not well understood.

The optical band gap is one of the most essential electronic properties of CP materials for their use in photonic and electronic devices, such as organic solar cells (OSCs), organic light-emitting diodes, and organic field-effect transistors.[14] Quantum chemistry simulations, particularly Density Functional Theory (DFT) and Time-Dependent DFT (TDDFT), are indispensable in polymer science for predicting and rationalizing the properties of polymeric materials.[15, 16] These methods offer insights into molecular properties across both ground and excited states and facilitate the prediction of optical gaps. While these simulations can handle sizeable molecular systems efficiently, the correlation between experimental measurements and DFT/TDDFT calculated values often remains weak for several reasons.[17, 18] The optical band gap is the energy required for a photon to excite an electron from the ground state to the first excited state, typically measured using UV-Vis absorption or photoluminescence spectroscopy. These measurements incorporate all physical effects presented in the system, such as solvent effects, vibronic coupling, and other fine details of the electronic structure. In contrast, the DFT-calculated HOMO-LUMO gap measures the energy required to move an electron from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO), ignoring the Coulombic interactions between the excited electron and the hole, which are involved in the optical band gap. Also, the HOMO-LUMO gap represents a vertical electronic transition and misses the relaxation transition due to interactions of the excited states with the surrounding environment. On the other hand, the TDDFT method calculates excited state energies based on the time-dependent response of the electronic system to an external perturbation. The lowest excited state in TDDFT corresponds to the energy required to promote an electron from the ground state to the first excited state, including excitonic effects and

accounting for the dynamic response of the electrons. This provides a more accurate description of excitation energies compared to static DFT. The accuracy of both DFT and TDDFT depends on the choice of exchange-correlation (xc) functional, with hybrid functionals generally yielding better results. However, discrepancies can still arise due to functional approximations and the lack of consideration for experimental conditions. Moreover, higher-level quantum mechanical theories, such as GW method coupled with Bethe-Salpeter equation, which might offer improved accuracy, are impractical for large systems such as CPs due to their computational demands.[19]

Data-driven machine learning (ML) approaches are powerful tools for the rapid property predictions and virtual structure screening of organic molecules and CP materials, offering substantial time and cost advantages over traditional experimental and computational methods.[17, 20–22] The effectiveness of ML models depends crucially on the availability of adequate and reliable training data, as well as the selection of appropriate descriptors that capture the structural and physicochemical properties of CPs. These descriptors include topological, electronic, geometrical, and molecular fragment attributes.[23, 24] Previous studies indicate that different ML algorithms trained with identical descriptors often yield similar accuracy. However, descriptor selection significantly impacts model performance, underscoring its dominant role in determining prediction accuracy.[25] So far, ML methods are increasingly used to predict the electronic properties of CPs and their derivatives, with most studies focusing on small molecules. The unique complexities of CP systems remain less explored, resulting in a scarcity of robust models that accurately reflect the behavior of these polymers. Additionally, many studies utilize DFT-calculated HOMO-LUMO gaps (E_{gap}^{DFT}) as reference data, which are generally less correlated with experimentally measured optical gaps (E_{gap}^{exp}) of CPs. Some studies incorporate E_{gap}^{exp} data from small CP datasets,[26] but these models generally show low performance and debatable robustness and transferability. Furthermore, due to the absence of underlying physical principles and the use of obscure descriptors, even well-trained ML models for interpolation struggle with robust performance when extrapolating to new CP design spaces.[20] To the best of our knowledge, there remains a significant gap in the development of well-established ML models that can predict the experimentally measured optical gap of CPs with high accuracy and transferability.

In this study, we developed a sophisticated approach that combines DFT calculations and data-driven ML models to accurately predict the E_{gap}^{exp} values of CPs. We demonstrated that by modifying oligomer structures—specifically, removing alkyl side chains and extending conjugated backbones—we can effectively capture the electronic properties of CPs. This modification significantly improved the correlation between E_{gap}^{DFT} and E_{gap}^{exp}, achieving an R^2 value of 0.51, while also considerably reducing computational time consumption. In contrast, unmodified monomer structures yielded a notably low R^2 value of 0.15. To further enhance the prediction accuracy of E_{gap}^{exp}, we trained a variety of ML models using both E_{gap}^{DFT} and conventional molecular representations as inputs. Compared to the baseline model trained with only molecular representations, incorporating E_{gap}^{DFT} of modified oligomers not only improves prediction accuracy—reflected by an R^2 of 0.77 and a mean absolute error (MAE) of 0.065 eV achieved by XGBoost—but also enhances the models' transferability in predicting the optical gaps of new polymers outside the design space of the training dataset. Our work outlines a rational strategy for predicting fundamental properties of polymers by segmenting

them into different substructures. These substructures are then characterized using different levels of theoretical or methodological approaches based on how well they correlate with the target properties. This methodological framework provides a robust basis for enhancing the predictive capabilities of computational models in polymer science.

2 Computational Details

2.1 Dataset

The original dataset with 1203 data points was adopted from Saeki et al's work,[27] in which experimentally measured data of synthesized polymers for OSC applications were manually collected from 503 literatures. For each polymer, the simplified molecular input line entry system (SMILES) string of its repeating unit was provided, together with a list of experimental parameters, including HOMO, LUMO, and E_{gap}^{exp}. We removed 88 duplicate entries based on the SMILES strings and 18 non-conjugated polymer structures containing sp^3-hybridized N atom along backbone chain (see Figure S1). An extra polymer containing Tellurium atoms was also excluded due to being out of the applicable range for the 6-31G* basis set in DFT calculations. Therefore, the final dataset comprised 1096 unique CPs. The distribution plots and statistical analysis of HOMO, LUMO, and E_{gap}^{exp} values are presented in Figure S2 and Table S1.

2.2 DFT calculations

All the DFT and TDDFT calculations were performed with Gaussian 16 package.[28] The B3LYP hybrid functional[29–31] together with D3 dispersion correction[32] and 6-31G* basis set were employed for both geometry optimization and electronic property calculations. The maximum force tolerance is 0.02 eV/Å. The initial xyz coordinates of polymers were generated from the SMILES strings with OpenBabel package.[33] Before geometry optimization with DFT, we manually adjusted the oligomer backbone to be coplanar using the Avogadro package[34] to more closely resemble a realistic configuration, as high planarity is favored in experiments to promote the performance of polymer-based electronic devices.[35]

2.3 Molecular Features

The chemical structures of CPs were represented with SMILES strings. RDKit library[36] was used to convert SMILES strings into three types of molecular features (MFs), including RDKit Descriptor,[37] molecular access system (MACCS),[38] and extended connectivity fingerprints (ECPF6).[39] RDKit Descriptor consists of the 209 molecular properties calculated by RDKit package, covering structural connectivity, geometry, electronic properties, and chemical composition. MACCS is a predefined fragment library with a subset of 166 keys which counts the presence of 166 various chemical fragments, such as S-N and alkaline metal, whereas one extra key with zero value is added as a consequence of Python's array-indexing-by-zero convention, resulting in a 167-bit vector. ECFP6, a flavor of Morgan fingerprints, considers the neighboring connectivity of atoms with 1024 keys, which

was generated by selecting the maximum diameter of the circular atom neighborhood to be six. We performed feature selection on 209 RDKit descriptors to eliminate irrelevant or redundant features, with the details presented in **Note S1**.

2.4 Machine Learning Models

We employed six conventional ML algorithms: Hist Gradient Boosting regression (HGBR),[40] Gradient Boosting Regression (GBR),[41] LightGBM regression (LGBM),[42] Extreme Gradient Boosting regression (XGBoost),[43] AdaBoost regression (AdaBoost),[44] random forest (RF).[45] These models are widely used in materials science and chemistry to uncover structure-property relationships.[46, 47] ML model training was performed with Scikit-learn library.[48] The details of the training process can be found in **Note S2**. Performance metrics, including coefficient of determination (R^2), Pearson correlation coefficient (r), Root mean square error (RMSE), and mean absolute error (MAE), were defined in **Note S3**.

3 Results and Discussion

3.1 Rational Design of Oligomer Model

CPs adopt a one-dimensional periodic structure, consisting of a conjugated backbone and various lengthy alkyl side chains. This periodicity and the lengthy side chains present significant challenges in modeling CP systems accurately and efficiently. CPs feature an extended backbone, but their poor crystallization results in an ill-defined lattice, complicating the construction of a periodic model for simulation. Furthermore, using a monomer—a single repeating unit—to represent a CP, fails to effectively capture the characteristics of π-electron delocalization in CP structures. Here, we employed monomer structures to calculate the E_{gap}^{DFT} values at the B3LYP level, denoted as $E_{gap}^{monomer}$. As shown in Figure 1b, $E_{gap}^{monomer}$ show no linear relationship with E_{gap}^{exp} of the corresponding CPs, as evidenced by a markedly low R^2 value of 0.15. This weak correlation is attributed to both the intrinsic limitations of the DFT method for predicting optical gaps[16, 49] and the inadequacy of monomer models in accurately representing the properties of CPs.

In this study, we rationally designed an oligomer model to represent CP materials based on their fundamental characteristics. The two-step procedure to construct the oligomer structure is depicted in Figure 1a with polymer PTB7 as an example. In the first step, we replaced the long alkyl side chains with methyl groups since alkyl side chains primarily affect solubility. This substitution enables efficient computational simulations while maintaining similar electronic and optical properties.[50, 51] As the conjugated backbone contributes most to the electronic properties of CPs, following side chain truncation, we replicated the monomer to form an oligomer structure, such as dimer or trimer. Previous studies suggested that using oligomer structures—such as a dimer, trimer, or tetramer, which extends the conjugated backbone chains to capture the characteristic of π-electron delocalization—enhances the predictive accuracy of DFT methods for experimental gaps, compared to employing monomer.[52] Given the limited understanding of the correlation between the conjugation length of backbone chain and the electronic properties of CPs, we tested various polymer

structures and established two guidelines for constructing oligomers (**see Note S4** for details): (1) the oligomer should contain at least four aromatic blocks linked by C-C single bonds along the backbone chain; (2) the oligomer should consist of at least six aromatic rings. Additionally, after side chain truncation, the obtained monomer can be regarded as an oligomer if it simultaneously contains more than four aromatic blocks and more than eight aromatic rings.

To validate the effectiveness of the simplified oligomer in capturing the electronic properties of CPs in comparison with the monomer, we also calculated the E_{gap}^{DFT} values with oligomers at the B3LYP level, denoted as $E_{gap}^{oligomer}$. Other xc functionals, including PBE,[53] ωB97XD,[54] and CAM-B3LYP,[55] exhibit high linear correlations with B3LYP for HOMO-LUMO gap calculations, achieving Pearson correlation coefficients above 0.96 among each other (**see Note S5**). As shown in Figure 1b, the modified oligomers exhibit a significant improvement in correlating DFT-calculated and experimental gap values compared to the monomers, with the R^2 value increasing from 0.15 to 0.51. This substantial enhancement underscores the importance of selecting appropriate configurations to accurately represent the fundamental characteristics of CP materials. On the other hand, our results suggest that ML models trained with DFT-calculated HOMO-LUMO gaps of monomers as reference data may not effectively predict experimental optical gaps of CPs due to the poor correlation observed.

Besides accuracy improvement, our two-step simplification procedure also significantly reduces computational demands by decreasing the number of atoms in CPs. Figure 1c illustrates the histogram distributions of atom counts for monomer and oligomer structures. Around 80% of the monomers, originally with atom counts ranging from 107 to 232, were reduced to between 78 and 156 atoms. Particularly, the largest monomer, which contains 494 atoms, is reduced to 200 atoms in its oligomer form, while the smallest with 33 atoms increases to 68 in the oligomer. This reduction is primarily achieved through the truncation of long alkyl side chains, while the increase in the number of atoms is due to the extension of the backbone chains. These two modifications synergistically lead to an overall decrease in system size for the majority of CPs.

CPs are composed of electron donor and acceptor units as building blocks linked by C-C single bonds. These donor and acceptor units are crucial for tuning the electronic properties, particularly the optical band gap.[56–58] Donor units donate electron density to the polymer backbone, raising the HOMO level, while acceptor units withdraw electron density, lowering the LUMO level. Thus, combining donor and acceptor units creates a push-pull effect, significantly narrowing the band gap and enhancing charge transport. By strategically incorporating donor and acceptor units into the polymer backbone, a variety of CPs can be designed with tailored electronic properties for specific applications. Given the importance of donor and acceptor units in determining the experimental optical gap, we aimed to investigate the effectiveness of our oligomer model in capturing the electronic properties of various donor and acceptor units. We categorized the 1096 CPs in our dataset based on donor or acceptor unit types. Figure 2c,f show the chemical structures of four commonly used donor and acceptor units. The categorization follows a specific order from D1 to D4, with polymers containing multiple donor types assigned to the latter type in the search order (e.g., D4 if containing both D1 and D4). D1 represents benzodithiophene and its derivatives with S atoms replaced by O and Se. D2, D3, and D4 represent carbazole, dithieno[3,2-b:2',3'-d]pyrrole,

and pyrroloindacenodithiophene, respectively, along with their derivatives where N is substituted by C, Si, O, S, and Se.[56] Polymers lacking these donor units are grouped as "Others". The same approach was applied for acceptor units. A1 represents the benzazole series, encompassing benzothiadiazole (BT), benzotriazole (BTA), benzoxazole, and related derivatives.[57] A2, A3, and A4 denote diketopyrrolo[3,4-c]-pyrrole-1,4-dione (DPP), quinoxaline (QA), and thieno[3,4-c]pyrrole-4,6-dione (TPD), respectively.[56] Based on the percentage distributions (see Table S8), D1 and A1 are the predominant donor and acceptor units, with ratios of 46.7% and 32.8%, respectively. As shown in Figure 2b,e, oligomers in each CP group exhibited significantly improved R^2 values compared to monomers (Figure 2a,d), reinforcing the widespread efficacy of our two-step approach in capturing the electronic properties of copolymers via modified oligomer structures, regardless of the specific donor and acceptor types.

3.2 Effect of Alkyl Side Chains

As demonstrated above, our two-step procedure for oligomer model construction through side chain truncation and backbone extension significantly improves the linear correlation between E_{gap}^{DFT} and E_{gap}^{exp}, increasing the R^2 value from 0.15 to 0.51 (see Figure 1b). Particularly, side chain truncation largely reduces computational cost, which is beneficial for high-throughput screening. Although less impactful on electronic properties than the conjugated backbone, alkyl side chains still require consideration in order to further improve the prediction accuracy of experimentally measured optical gaps. Consequently, we applied two categories of descriptors for ML modeling: DFT calculated HOMO-LUMO gaps of modified oligomers to represent the extended backbone, and molecular features from SMILES strings of monomers to capture the effect of alkyl side chains.

In this study, we evaluated the effectiveness of three types of MFs for capturing the impact of alkyl side chains on the optical gaps of CPs: RDKit Descriptors, MACCS, and ECFP6 fingerprints, which were calculated from the SMILES strings of monomer structures containing alkyl side chains as detailed in the Methods section. Previous studies have shown these MFs are effective for training ML models to predict the photo-electronic properties of CP-based OSCs.[27, 59] The workflow for database preparation, feature engineering, model training, and transferability test is summarized in Figure 3. We trained six ML algorithms that were commonly used in materials sciences, including HGBR, LGBM, GBR, XGBoost, AdaBoost, and RF, using E_{gap}^{DFT} combined with different types of MFs as input parameters to predict E_{gap}^{exp}. Performance metrics, including R^2, r, RMSE, and MAE, were detailed in Table S9-S11. Notably, the prediction accuracy of optical gap values is significantly enhanced, with the R^2 value increasing from 0.51 to as high as 0.77, by incorporating information from both the conjugated backbone (captured by $E_{gap}^{oligomer}$) and the side chains (captured by MFs). Particularly, ECFP6 combined with $E_{gap}^{oligomer}$ consistently achieved the highest prediction accuracy across all ML models, demonstrating its superiority in capturing the side chain information of CPs compared to RDKit and MACCS.

To further investigate the impact of $E_{gap}^{oligomer}$ on optical gap prediction, we summarized the R^2 and MAE values for the six ML models trained using ECFP6 alone and in combination with $E_{gap}^{oligomer}$ in Figure 4. All models achieved higher accuracy using $E_{gap}^{oligomer}$ and ECFP6, with R^2 values

over 0.62, compared to 0.51 for a simple linear regression model (Figure 1c). Notably, XGBoost emerged as the top performer with an R^2 of 0.77 and MAE of 0.065 eV. This level of accuracy falls within the experimental error margin of approximately 0.1 eV. For instance, polymer P3HT has an E_{gap}^{exp} between 1.9 and 2.14 eV,[60–63] while PDB7 ranges from 1.6 to 1.7 eV,[64–66] influenced by molecular weight, regioregularity, and processing conditions.[67] In addition, when retrained with only ECFP6, all models had lower accuracy; for example, the XBGoost model achieved an R^2 of 0.7 and MAE of 0.075 eV.

In summary, descriptors are essential for ML models to capture critical information influencing targeted properties and learn structure-property relationships effectively. DFT-calculated HOMO-LUMO gaps of modified oligomers and ECFP6 MF derived from unmodified monomers can effectively capture fundamental characteristics of both the extended backbone and alkyl side chains, enabling accurate and efficient prediction of E_{gap}^{exp} values.

3.3 Model Transferability

We have demonstrated that ML models can leverage rationally designed $E_{gap}^{oligomer}$ and MFs to improve the prediction accuracy of experimental optical gaps. Beyond accuracy, it is essential to validate the model's robustness and transferability with new datasets which have not been used in the ML model training process. In this study, we manually collected 227 newly synthesized CP structures from the literature, categorizing them into two groups based on their electron acceptor units. As shown in Figure 5a, CP structures from group 1 contain at least one of the five acceptor units which were included in the training set; for example, BT and BTA units belong to A1 type (see Figure 2f). This subset of CP structures is applied to evaluate the interpolation performance of the trained ML models considering the close similarity of this subset of CP structures as compared to the training set. In contrast, group 2 contains five acceptor units which have not been seen in the training set (see Figure 5b), including Perylene Diimide (PDI),[68] naphthalenediimide (NDI),[69] dithieno[3',2':3,4;2",3":5,6]benzo[1,2-c][1,2,5]thiadiazole (DTBT),[70] benzobisoxazole (BBX),[71] and Y6.[72] These acceptor units are important components of electron-accepting semiconductors for organic photovoltaic applications. For example, the Y6-based small molecule, first announced in 2019, achieved a record power conversion efficiency of 15.7% as an acceptor in OSCs.[73] In 2020, the first Y6-series-based polymer acceptor was reported, and since then, these acceptors have been recognized as the best n-type materials.[74] Additionally, Ding et al. introduced a novel family of polymer donors named D18 based on DTBT and fluorinated BDTT in 2020, achieving the first single-junction OSC with an efficiency of over 18% when blended with Y6 small molecule.[75] The CP structures from group 2 containing one of these five acceptor units are used to assess the extrapolation performance of the trained ML models.

For both group 1 and group 2 CP structures, we constructed the modified oligomers using the two-step procedure (see Figure 1a) to obtain the $E_{gap}^{oligomer}$ values and converted the SMILES strings of the alkyl-side-chain-containing monomers into ECFP6 features. Then, we applied the XGBoost models previously trained with the 1096 dataset to predict the E_{gap}^{exp} of both groups without further retraining. The performance metrics are presented in Table 1. XGBoost-2, trained with $E_{gap}^{oligomer}$

and ECFP6, accurately predicted the optical gaps of group 1 CPs with most MAEs below 0.1 eV, demonstrating excellent interpolation performance. Interestingly, XGBoost-1, trained with only ECFP6, also showed superior interpolation performance, resulting in lower RMSE and MAE than XGBoost-2 across all CP types in group 1. These results suggest that ECFP6 effectively captures the electronic properties of similar CPs within the same chemical design space. In fact, ECFP6 has been widely used to measure the similarities of various organic molecules in previous studies.[76]

We then assessed the extrapolation performance of both models with group 2 CPs. As shown in Table 1, XGBoost-2 significantly outperformed XGBoost-1, yielding substantially lower RMSE and MAE values across all CP types. For example, the MAE for Y6 based CPs decreases from 0.418 eV to 0.211 eV with XGBoost-2. Previous studies have also shown that conventional ML models trained with molecular descriptors may perform well in the chemical structure space similar to the training set, whereas the extrapolation to the new structure space is challenging due to the lack of physical/chemical insight from the input descriptors.[77, 78] Our results demonstrate that the XGBoost-2 model, trained with both $E_{gap}^{oligomer}$ and ECFP6, excels in both high interpolation and extrapolation performance. This superior transferability is originated from the excellent robustness of DFT methods and rationally designed oligomer structures, effectively capturing the electronic properties of the CPs. Indeed, as shown in Figure 5c,d, $E_{gap}^{oligomer}$ are highly correlated with E_{gap}^{exp} for both group 1 and group 2 CPs.

3.4 Further Discussion

As detailed above, the XGBoost-2 model, trained with 1096 CPs, demonstrated the superior transferability on a new dataset of 227 CPs. It is well acknowledged that conventional ML models such as XGBoost can benefit from larger datasets to further improve prediction accuracy.[79, 80] Therefore, we selected one structure with the highest prediction error from each category in group 1 and group 2, forming a new test set of 10 CP structures (see Figure S7). The remaining 217 data points were combined with the original 1096, creating a new training set of 1313 structures, thus augmenting the training set by around 20%. We retrained a new XBGoost model (labeled as "XGBoost-2-plus") with 10-fold cross-validation on 1313 data points and calculated the average RMSE and MAE for predicting the E_{gap}^{exp} of 10 CPs in the new test set. As shown in Table S12, XGBoost-2-plus achieved enhanced prediction accuracy, with a lower RMSE of 0.241 eV and MAE of 0.213 eV compared to XGBoost-2 (0.333 eV and 0.3 eV, respectively). Particularly, the prediction errors of the XGBoost-2-plus were significantly reduced for each polymer in the test set, demonstrating the effectiveness of data augmentation in improving model performance. It is important to note that experimentally measured optical gap values for CPs can vary across different labs and experiments, introducing potential inconsistencies and errors. Factors such as processing conditions, solvents, additives, and film morphology can influence these measurements.[81] Utilizing larger and more accurate experimental datasets can enhance the predictive accuracy of ML models for optical gaps.

In addition to predicting optical gaps, our training strategy, which combines quantum chemistry calculations and MFs, extends to predicting other fundamental properties of CPs such as HOMO and LUMO levels, which are also vital for their applications in electronics and solar cells.[82, 83] In our

dataset of 1096 structures, HOMO levels were measured via cyclic voltammetry, while LUMO levels were derived from substituting optical gap values with HOMO values. Following the methodology detailed in the Methods section, we retrained the XGBoost model and presented the performance metrics in Table S13. Notably, XGBoost-2-H trained with DFT-calculated HOMO values of modified oligomers and ECFP6 exhibits higher accuracy in predicting experimentally measured HOMO values, achieving an R^2 of 0.5 and an MAE of 0.109 eV. Similarly, incorporating DFT-calculated LUMO levels enhances the accuracy of XGBoost-2 in predicting LUMO levels, with an R^2 of 0.6 and an MAE of 0.112 eV.

Conclusions

In this study, we introduced a model that combines DFT calculations with a ML approach to accurately predict the experimentally measured optical band gaps of CPs, utilizing a dataset of 1096 data points. We first proposed a two-step modification procedure for constructing oligomers to effectively capture π-electron delocalization in CPs: alkyl side chain truncation and conjugated backbone extension. This approach significantly improves the correlation between the DFT-calculated HOMO-LUMO gaps and experimental gaps (R^2=0.51) compared to the unmodified side-chain-containing monomers (R^2=0.15). Subsequently, we incorporated both conjugated backbone characteristics, derived from quantum chemistry, and the alkyl-side-chain effects, represented by molecular descriptors, into ML modeling to enhance prediction accuracy. Employing the $E_{gap}^{oligomer}$ of modified oligomers and ECFP6 MF derived from side-chain-containing monomers as input, the resulting model, XGBoost-2, effectively elucidated the structure-property relationship of CPs, achieving an R^2 of 0.77 and an MAE of 0.065 eV. To further assess its robustness and transferability in predicting new CP structures beyond the chemical design space of the training set, we manually collected 227 newly synthesized CPs from the literature, categorizing them into two groups based on their electron acceptor units. Group 1 CP structures contain at least one of the five acceptor units existing in the training set, allowing for the evaluation of interpolation performance, while Group 2 structures contain at least one of the five acceptor units not present in the training set, aiming for extrapolation performance test. Notably, XGBoost-2 demonstrates excellent interpolation and extrapolation, which stem from the combination of DFT methods and rationally designed oligomer structures that effectively capture the electronic properties of CPs. This study represents the first successful combination of quantum chemistry calculations with ML modeling to accurately predict experimentally measured fundamental properties of CPs (e.g., HOMO, LUMO, and optical gap), facilitating the design and development of next-generation high-performance CPs in photoelectronic and energy conversion applications.

Author contributions

B.L. and M.L. initiated this study. B.L. was responsible for conducting the theoretical calculations, training the machine learning models, and analyzing the data. Y.Y. was responsible for data collection, structural generation, and optimization. All authors contributed to the manuscript writing.

Acknowledgements

This work was financially supported by the University of Florida's new faculty start-up funding. The authors acknowledge the University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used and/or analysed during the current study are available from:https://github.com/Liu-Group-UF/Machine-Learning-for-Accurate-Optical-Gap-Prediction-in-Conjugated-Polymers

Code availability

The Python code used for Machine Learning model training is available from:https://github.com/Liu-Group-UF/Machine-Learning-for-Accurate-Optical-Gap-Prediction-in-Conjugated-Polymers

Supporting Information

The Supporting Information is available free of charge.

- RDKit Descriptor Selection. Model Training Strategy. Model Performance Metrics. Oligomer structure construction. Exchange-correlation functional test. Statistical analysis of the experimentally measured HOMO, LUMO, and optical band gap values. Chemical structures of 18 non-conjugated polymers. The number of data points and the corresponding percentage of each group of conjugated polymers categorized based on donor and acceptor units. The performance metrics of 6 Machine Learning models.

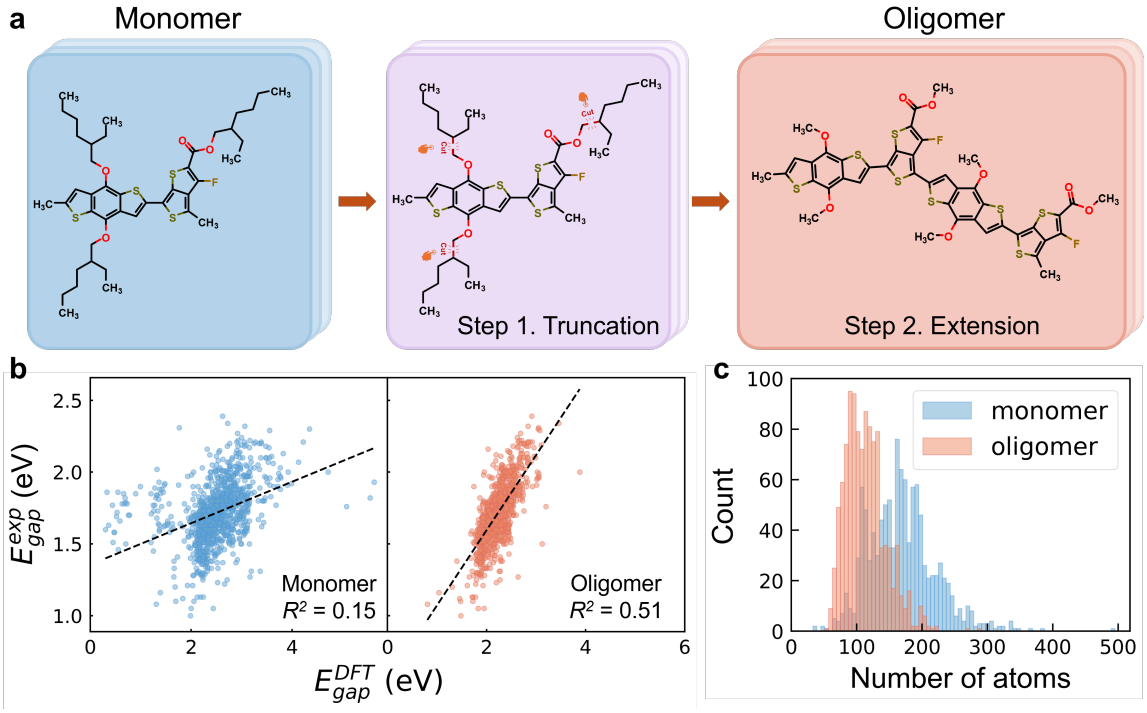

Figure 1: (a) Scheme of converting the monomer structure of PTB7 into a modified oligomer with a two-step procedure, namely, alkyl side chain truncation and conjugated backbone extension. (b) The parity plots of DFT calculated HOMO-LUMO gaps (E_{gap}^{DFT}) based on monomer and modified oligomer structures versus experimentally measured optical gaps (E_{gap}^{exp}). The black dashed lines correspond to linear fitting. (c) The distributions of atom counts in the monomers and modified oligomers.

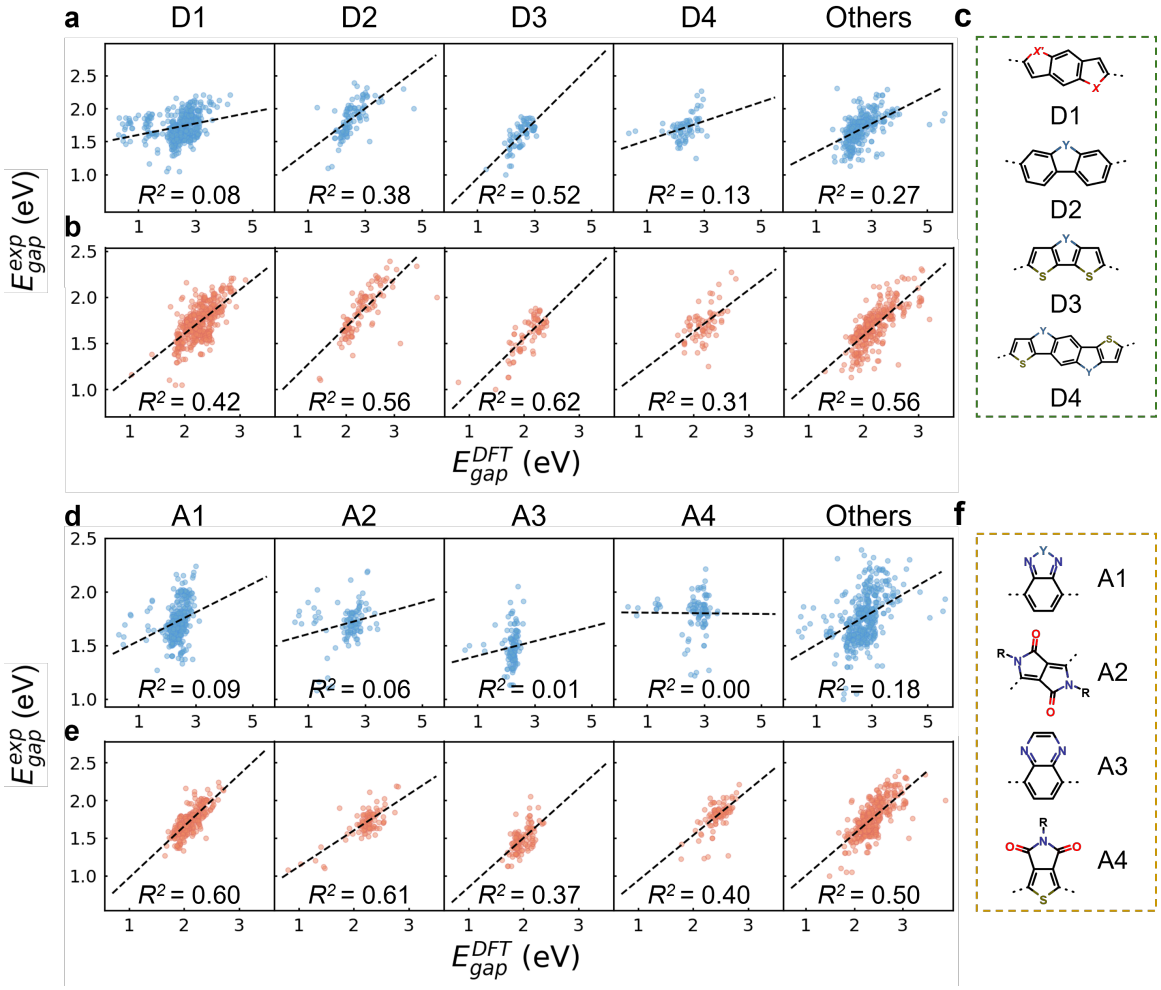

Figure 2: The linear correlation between DFT-calculated HOMO-LUMO gaps (E_{gap}^{DFT}) and experimental optical gaps (E_{gap}^{exp}) for different groups of conjugated polymers categorized based on donor and acceptor units, respectively. The E_{gap}^{DFT} is calculated from (a,d) monomers with alkyl side chains and (b,e) modified oligomers after two-step procedure shown in Figure 1a. The black dashed lines correspond to linear fitting. (c,f) The chemical structures of four donor and acceptor units. X or X' denotes O, S, or Se atom, and Y represents C, N, O, Si, S, or Se atom.

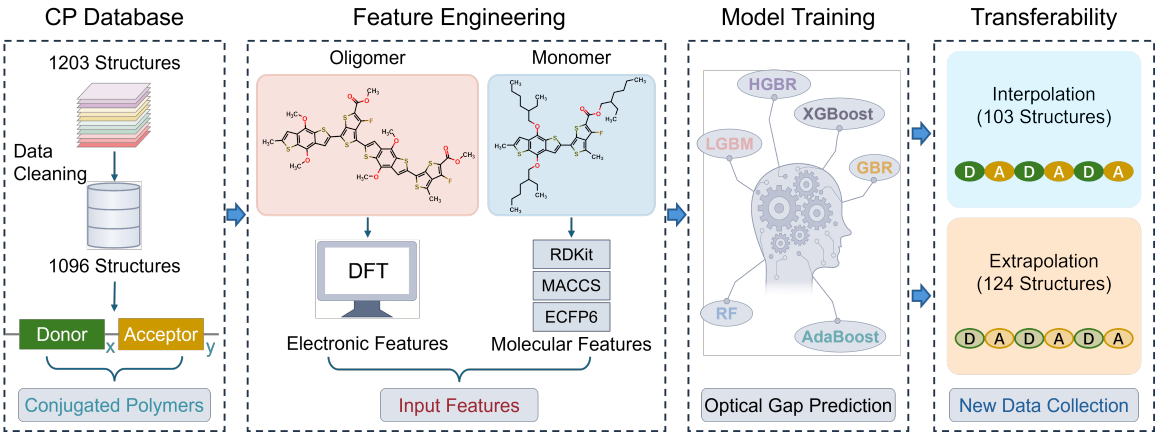

Figure 3: The workflow for the machine learning model training procedure to predict the experimentally measured optical gaps of conjugated polymers (CPs).

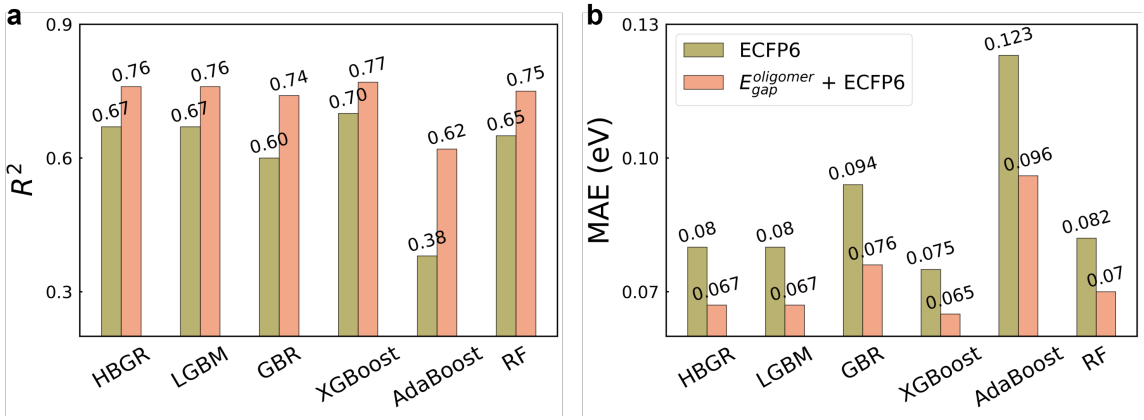

Figure 4: (a) R^2 and (b) mean absolute error (MAE) (eV) of six machine learning models for predicting experimental optical gaps of conjugated polymers with different descriptors as input. $E_g^{oligomer}$ is HOMO-LUMO gap calculated from modified oligomer structures.

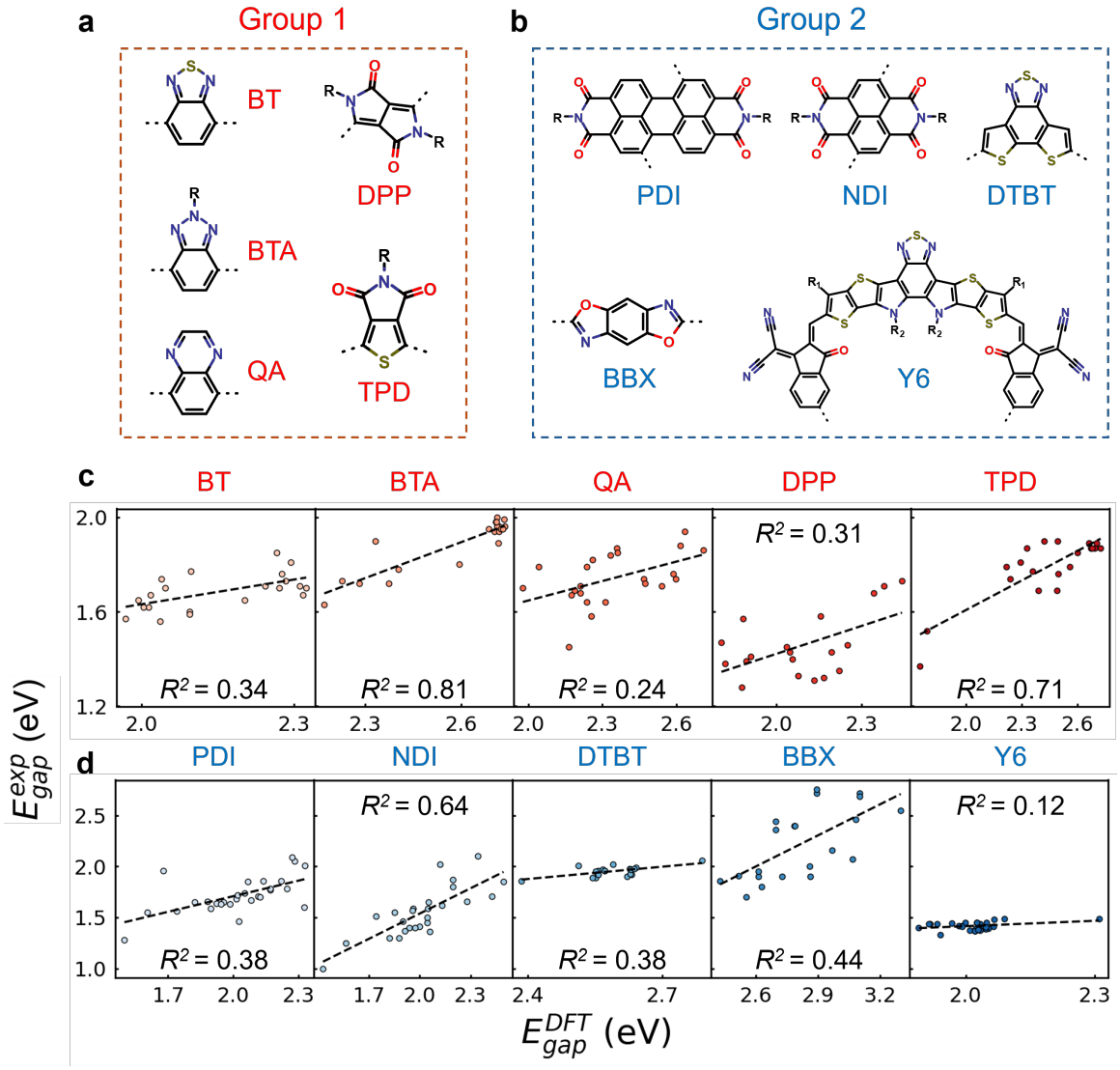

Figure 5: The chemical structures of (a) five acceptor units exiting in the training set and (b) five acceptor units not existing in the training set. (c,d) The linear correlation between DFT-calculated HOMO-LUMO gaps (E_{gap}^{DFT}) with modified oligomer structures and experimental optical gaps (E_{gap}^{exp}) for (c) group 1 and (d) group 2 conjugated polymers. The black dashed lines correspond to linear fitting.

Table 1: The performance metrics of XGBoost-1 and XGBoost-2 in predicting the experimental optical gaps of conjugated polymers categorized by various acceptor units. Group 1 acceptor units are included in the training set, whereas Group 2 units are not. Chemical structures are illustrated in Figure 5. XGBoost-1 is trained using ECFP6 alone, while XGBoost-2 is trained with both $E_{gap}^{oligomer}$ and ECFP6. Both root mean square error (RMSE) and mean absolute error (MAE) are measured in eV.

		XGBoost-1		XGBoost-2	
Acceptor unit	#Data points	RMSE	MAE	RMSE	MAE
Group 1					
BT	21	0.073	0.052	0.085	0.069
BTA	20	0.106	0.085	0.112	0.099
QA	23	0.075	0.063	0.114	0.092
DPP	19	0.068	0.054	0.133	0.102
TPD	20	0.077	0.063	0.09	0.068
Group 2					
PDI	28	0.181	0.158	0.189	0.147
NDI	26	0.262	0.211	0.126	0.094
DTBT	20	0.157	0.135	0.059	0.041
BBX	21	0.583	0.49	0.338	0.253
Y6	29	0.427	0.418	0.217	0.211

References

1. Palani P and Karpagam S. Conjugated polymers–a versatile platform for various photophysical, electrochemical and biomedical applications: a comprehensive review. New Journal of Chemistry 2021;45:19182–209.
2. Malik AH, Habib F, Qazi MJ, Ganayee MA, Ahmad Z, and Yatoo MA. A short review article on conjugated polymers. Journal of Polymer Research 2023;30:115.
3. Kang S, Yoon TW, Kim GY, and Kang B. Review of conjugated polymer nanoparticles: from formulation to applications. ACS Applied Nano Materials 2022;5:17436–60.
4. Lu L, Zheng T, Wu Q, Schneider AM, Zhao D, and Yu L. Recent advances in bulk heterojunction polymer solar cells. Chemical reviews 2015;115:12666–731.
5. Liu C, Wang K, Gong X, and Heeger AJ. Low bandgap semiconducting polymers for polymeric photovoltaics. Chemical Society Reviews 2016;45:4825–46.
6. Li Y, Jia Z, Zhang Q, et al. Toward efficient all-polymer solar cells via halogenation on polymer acceptors. ACS applied materials & interfaces 2020;12:33028–38.
7. Pace G, Bargigia I, Noh YY, Silva C, and Caironi M. Intrinsically distinct hole and electron transport in conjugated polymers controlled by intra and intermolecular interactions. Nature communications 2019;10:5226.
8. Liu B, Rocca D, Yan H, and Pan D. Beyond conformational control: effects of noncovalent interactions on molecular electronic properties of conjugated polymers. Jacs Au 2021;1:2182–7.
9. Bin H, Yang Y, Peng Z, et al. Effect of Alkylsilyl Side-Chain Structure on Photovoltaic Properties of Conjugated Polymer Donors. Advanced Energy Materials 2018;8:1702324.
10. Chen X, Hussain S, Hao Y, Tian X, and Gao R. Recent advances of signal amplified smart conjugated polymers for optical detection on solid support. ECS Journal of Solid State Science and Technology 2021;10:037006.
11. Liu Y, Feig VR, and Bao Z. Conjugated polymer for implantable electronics toward clinical application. Advanced Healthcare Materials 2021;10:2001916.
12. Luong JH, Narayan T, Solanki S, and Malhotra BD. Recent advances of conducting polymers and their composites for electrochemical biosensing applications. Journal of Functional Biomaterials 2020;11:71.
13. Zhao C, Chen Z, Shi R, Yang X, and Zhang T. Recent advances in conjugated polymers for visible-light-driven water splitting. Advanced Materials 2020;32:1907296.
14. Scharber MC and Sariciftci NS. Low band gap conjugated semiconducting polymers. Advanced Materials Technologies 2021;6:2000857.
15. Laurent AD and Jacquemin D. TD-DFT benchmarks: a review. International Journal of Quantum Chemistry 2013;113:2019–39.

16. Sun H and Autschbach J. Electronic energy gaps for π-conjugated oligomers and polymers calculated with density functional theory. Journal of Chemical Theory and Computation 2014;10:1035–47.

17. Pyzer-Knapp EO, Simm GN, and Guzik AA. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. Materials Horizons 2016;3:226–33.

18. Yang SJ, Li S, Venugopalan S, et al. Accurate Prediction of Experimental Band Gaps from Large Language Model-Based Data Extraction. arXiv preprint arXiv:2311.13778 2023.

19. Chaudhuri D and Patterson C. TDDFT versus GW/BSE Methods for Prediction of Light Absorption and Emission in a TADF Emitter. The Journal of Physical Chemistry A 2022;126:9627–43.

20. Meftahi N, Klymenko M, Christofferson AJ, Bach U, Winkler DA, and Russo SP. Machine learning property prediction for organic photovoltaic devices. npj computational materials 2020;6:166.

21. Pyzer-Knapp EO, Li K, and Aspuru-Guzik A. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. Advanced Functional Materials 2015;25:6495–502.

22. Mazouin B, Schöpfer AA, and Lilienfeld OA von. Selected machine learning of HOMO–LUMO gaps with improved data-efficiency. Materials Advances 2022;3:8306–16.

23. McGibbon M, Shave S, Dong J, et al. From intuition to AI: evolution of small molecule representations in drug discovery. Briefings in bioinformatics 2024;25:bbad422.

24. Raghunathan S and Priyakumar UD. Molecular representations for machine learning applications in chemistry. International Journal of Quantum Chemistry 2022;122:e26870.

25. Winkler DA and Le TC. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. Molecular informatics 2017;36:1600118.

26. Wu K, Sukumar N, Lanzillo N, et al. Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. Journal of Polymer Science Part B: Polymer Physics 2016;54:2082–91.

27. Nagasawa S, Al-Naamani E, and Saeki A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. The Journal of Physical Chemistry Letters 2018;9:2639–46.

28. Frisch M, Trucks G, Schlegel H, et al. GAUSSIAN16. Revision C. 01. Gaussian Inc., Wallingford, CT, USA. 2016.

29. Becke AD. A new mixing of Hartree–Fock and local density-functional theories. The Journal of chemical physics 1993;98:1372–7.

30. Stephens PJ, Devlin FJ, Chabalowski CF, and Frisch MJ. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. The Journal of physical chemistry 1994;98:11623–7.

31. Lee C, Yang W, and Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Physical review B 1988;37:785.

32. Grimme S, Antony J, Ehrlich S, and Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. The Journal of chemical physics 2010;132:154104.

33. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, and Hutchison GR. Open Babel: An open chemical toolbox. Journal of cheminformatics 2011;3:1–14.

34. Avogadro. URL: `https://avogadro.cc/`.

35. Goldey MB, Reid D, Pablo J de, and Galli G. Planarity and multiple components promote organic photovoltaic efficiency by improving electronic transport. Physical Chemistry Chemical Physics 2016;18:31388–99.

36. RDKit: Open-Source Cheminformatics Software. URL: `https://www.rdkit.org/`.

37. RDKit Descriptors module. URL: `https://rdkit.org/docs/source/rdkit.Chem.Descriptors.html`.

38. MACCSkeys module. URL: `http://rdkit.org/docs/source/rdkit.Chem.MACCSkeys.html`.

39. Rogers D and Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling 2010;50:742–54.

40. Guryanov A. Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In: *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers 8*. Springer. 2019:39–50.

41. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis 2002;38:367–78.

42. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017;30:1–9.

43. Chen T and Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016:785–94.

44. Solomatine DP and Shrestha DL. AdaBoost. RT: a boosting algorithm for regression problems. In: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*. Vol. 2. IEEE. 2004:1163–8.

45. Liaw A, Wiener M, et al. Classification and regression by randomForest. R news 2002;2:18–22.

46. Wei J, Chu X, Sun XY, et al. Machine learning in materials science. InfoMat 2019;1:338–58.

47. Morgan D and Jacobs R. Opportunities and challenges for machine learning in materials science. Annual Review of Materials Research 2020;50:71–103.

48. scikit-learn: Machine Learning in Python. URL: `https://scikit-learn.org/stable/`.

49. Bredas JL. Mind the gap! Materials Horizons 2014;1:17–9.

50. Falke SM, Rozzi CA, Brida D, et al. Coherent ultrafast charge transfer in an organic photovoltaic blend. Science 2014;344:1001–5.

51. Liu B, Chow PC, Liu J, and Pan D. Polarized local excitons assist charge dissociation in Y6-based nonfullerene organic solar cells: a nonadiabatic molecular dynamics study. Journal of Materials Chemistry A 2024.

52. Larsen RE. Simple extrapolation method to predict the electronic structure of conjugated polymers from calculations on oligomers. The Journal of Physical Chemistry C 2016;120:9650–60.

53. Perdew JP, Burke K, and Ernzerhof M. Generalized gradient approximation made simple. Physical review letters 1996;77:3865.

54. Chai JD and Head-Gordon M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. Physical Chemistry Chemical Physics 2008;10:6615–20.

55. Yanai T, Tew DP, and Handy NC. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). Chemical physics letters 2004;393:51–7.

56. Zhang ZG and Wang J. Structures and properties of conjugated donor–acceptor copolymers for solar cell applications. Journal of Materials Chemistry 2012;22:4178–87.

57. Chua MH, Zhu Q, Tang T, Shah KW, and Xu J. Diversity of electron acceptor groups in donor–acceptor type electrochromic conjugated polymers. Solar Energy Materials and Solar Cells 2019;197:32–75.

58. Hildner R, Köhler A, Müller-Buschbaum P, Panzer F, and Thelakkat M. π-Conjugated Donor Polymers: Structure Formation and Morphology in Solution, Bulk and Photovoltaic Blends. Advanced energy materials 2017;7:1700314.

59. Sun W, Zheng Y, Yang K, et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science advances 2019;5:eaay4275.

60. Kesornsit S, Direksilp C, Phasuksom K, et al. Synthesis of highly conductive poly (3-hexylthiophene) by chemical oxidative polymerization using surfactant templates. Polymers 2022;14:3860.

61. Aruna P and Joseph C. Optical and photosensing properties of gold nanoparticles doped poly (3-hexylthiophene-2, 5-diyl) thin films. Materials Letters 2021;295:129726.

62. Jung IH, Hong CT, Lee UH, Kang YH, Jang KS, and Cho SY. High thermoelectric power factor of a diketopyrrolopyrrole-based low bandgap polymer via finely tuned doping engineering. Scientific Reports 2017;7:44704.

63. Prasad SS, Divya G, and Kumar KS. P3HT Thin Films And Their Optical Characterization. In: *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. IEEE. 2021:5–8.

64. Liang Y, Xu Z, Xia J, et al. For the bright future-bulk heterojunction polymer solar cells with power conversion efficiency of 7.4%. Advanced materials 2010;22:E135.

65. Bencheikh F, Duché D, Ruiz CM, Simon JJ, and Escoubas L. Study of optical properties and molecular aggregation of conjugated low band gap copolymers: PTB7 and PTB7-Th. The Journal of Physical Chemistry C 2015;119:24643–8.

66. Basel T, Huynh U, Zheng T, Xu T, Yu L, and Vardeny ZV. Optical, electrical, and magnetic studies of organic solar cells based on low bandgap copolymer with spin 1/2 radical additives. Advanced Functional Materials 2015;25:1895–902.

67. He Y, Huo L, and Zheng B. Advances of batch-variation control for photovoltaic polymers. Nano Energy 2024:109397.

68. Shi Q, Wu J, Wu X, Peng A, and Huang H. Perylene Diimide-Based Conjugated Polymers for All-Polymer Solar Cells. Chemistry–A European Journal 2020;26:12510–22.

69. Zhou N and Facchetti A. Naphthalenediimide (NDI) polymers for all-polymer photovoltaics. Materials Today 2018;21:377–90.

70. Cao J, Yi L, Zhang L, Zou Y, and Ding L. Wide-bandgap polymer donors for non-fullerene organic solar cells. Journal of Materials Chemistry A 2023;11:17–30.

71. Intemann JJ, Hellerich ES, Ewan MD, et al. Investigating the impact of conjugation pathway on the physical and electronic properties of benzobisoxazole-containing polymers. Journal of Materials Chemistry C 2017;5:12839–47.

72. Kataria M, Chau HD, Kwon NY, Park SH, Cho MJ, and Choi DH. Y-series-based polymer acceptors for high-performance all-polymer solar cells in binary and non-binary systems. ACS Energy Letters 2022;7:3835–54.

73. Yuan J, Zhang Y, Zhou L, et al. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core. Joule 2019;3:1140–51.

74. Jia T, Zhang J, Zhong W, et al. 14.4% efficiency all-polymer solar cell with broad absorption and low energy loss enabled by a novel polymer acceptor. Nano Energy 2020;72:104718.

75. Liu Q, Jiang Y, Jin K, et al. 18% Efficiency organic solar cells. Science Bulletin 2020;65:272–5.

76. Riniker S and Landrum GA. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. Journal of cheminformatics 2013;5:1–7.

77. Zhao ZW, Del Cueto M, and Troisi A. Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors. Digital Discovery 2022;1:266–76.

78. Muckley ES, Saal JE, Meredig B, Roper CS, and Martin JH. Interpretable models for extrapolation in scientific machine learning. Digital Discovery 2023;2:1425–35.

79. Jin J, Faraji S, Liu B, and Liu M. Comparative Analysis of Conventional Machine Learning and Graph Neural Network Models for Perovskite Property Prediction. ChemRxiv 2024.

80. Ramezan CA, Warner TA, Maxwell AE, and Price BS. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. Remote Sensing 2021;13:368.

81. Fischer F, Tremel K, Saur AK, et al. Influence of processing solvents on optical properties and morphology of a semicrystalline low bandgap polymer in the neutral and charged states. Macromolecules 2013;46:4924–31.

82. Li Y. Molecular design of photovoltaic materials for polymer solar cells: toward suitable electronic energy levels and broad absorption. Accounts of chemical research 2012;45:723–33.

83. Jung JW, Jo JW, Jung EH, and Jo WH. Recent progress in high efficiency polymer solar cells by rational design and energy level tuning of low bandgap copolymers with various electron-withdrawing units. Organic Electronics 2016;31:149–70.