PAPER

Statistical Prediction of Donor-Acceptor Thiophene Copolymer Properties

Michael T. Cole,[a] Ilana Y. Kanal[a] and Geoffrey R. Hutchison*[a]

Conjugated polymers (CPs) have proven to be useful materials in organic optoelectronic devices. Polythiophenes are particularly interesting CPs due to the tunable nature of their electronic structures and their relative environmental stability. In this computational study, 91 diversely substituted thiophene monomers were combined into 4095 unique dimer pairs. These dimers were used to construct alternating $(AB)_n$ oligomers (4095 tetramers, 120 hexamers, and 70 octamers). DFT-B3LYP/6-31G* geometry optimizations were performed and the HOMO and LUMO energies were extracted for each of these 8471 chemical structures. Infinite polymer HOMO, LUMO, and HUMO/LUMO gap (E_g) energies were extrapolated for the 70 AB combinations for which hexamer and octamer calculations had been performed. Statistical models were developed to relate monomer and small oligomer electronic properties to the extrapolated polymer energies. It was found that monomer data alone was not sufficient to predict the extrapolated electronic energies of the infinite polymers using our statistical models. The inclusion of tetramer and dimer data into these models provided a significant increase in robustness of their predictive power.

Design, System, Application

This work illustrates some of the shortcomings of the standard donor-acceptor model as a tool for predicting polymer molecular orbital interactions. 8741 DFT calculations were performed for a series of co-oligomers of varying length from a diverse set of 91 thiophene monomers to explore statistical relationships between the frontier molecular orbital energies of oligomers and their degree of polymerization. These relationships were used to develop predictive models that allow for the calculation of polymer frontier molecular orbital energies. Polymer frontier molecular orbital energies have been shown to impact the device performances of many types of optoelectronic devices, including organic field effect transistors, organic photovoltaics and organic light emitting diodes.

Introduction

In recent years, conjugated polymers (CPs) have been of substantial research interest as new materials for use in a variety of optoelectronic devices, such as organic light-emitting diodes (OLEDs), organic field-effect transistors (OFETs), and organic photovoltaics (OPVs)[1,2]. CPs have several advantages over conventional inorganic materials including their relatively low cost and roll-to-roll processability[3]. Polythiophenes have been widely studied in semiconductor applications due to the tunability of their electronic properties via substitution[4,5] and sequence control[6,7], as well as their environmental stability[8,9]. It has been shown that OLED, OFET, and OPV device performances are related to the frontier molecular orbital (FMO) energies of the conjugated polymers from which they are composed[1,2,10]. This dependence of device performance on polymer FMO energies necessitates an efficient methodology for predicting these energies to streamline the discovery of new compounds for use in next generation optoelectronic devices.

The available chemical space that contains conjugated monomers of possible interest to polymer optoelectronic applications is on the order of 10^{60} molecules[11]. To complicate matters further, conjugated monomers can be combined into random (A_xB_y), alternating $(AB)_n$, or block $(A)_x(B)_y$ copolymers. Terpolymers and quaterpolymers of varying complexity can also be synthesized. Previous work has shown that changes in copolymer sequence for small oligomers has been shown to have significant effects on frontier molecular orbital energies[12]. To search this vast chemical space efficiently for molecules of interest, computational methods must be employed.

$$\begin{pmatrix} M_1M_2 & M_1M_3 & \cdots & M_1M_{91} \\ & M_2M_3 & \cdots & M_2M_{91} \\ & & \ddots & \vdots \\ & & & M_{90}M_{91} \end{pmatrix}$$

Figure 1. Monomer combination scheme for co-dimers

Calculations performed with density functional theory (DFT) and a hybrid functional such as B3LYP are typically in good agreement with experimental results for CP systems, although this agreement with experiment is often attributed to fortuitous error cancellations[13]. DFT calculations have a fast basis set convergence, and gas phase B3LYP/6-31G* calculations offer a good compromise between accuracy and calculation time. HOMO, LUMO, and HOMO/LUMO gap (E_g) energies extracted from DFT calculations are related to the ionization potential, electron affinity, and bandgaps respectively of conjugated organic molecules[13,14].

It is generally accepted that a linear relationship exists between the HOMO, LUMO, and the E_g of conjugated oligomers

and the inverse number of repeat units (1/n) in the molecule[13, 15-17]. This relationship can be used to extrapolate the HOMO/LUMO data for an infinite conjugated polymer from its short-chain oligomers. It is now known that these extrapolations deviate from linearity at large values of n. This is due to the saturation effect, or the maximum number of repeat units that can contribute to π delocalization along the polymer backbone. The saturation effect depends strongly on the conformation of the polymer, which can be adversely affected by monomer substitutions, solvent effects, and backbone rigidity. Many models have been developed to account for saturation effects[16], and terms such as the effective conjugation length[18] and the maximum conductive chain length[19] have been established to describe and quantify this phenomenon. When utilizing the simple 1/n extrapolation method to determine the E_g of a polymer from its constituent oligomers, the predicted infinite polymer HOMO/LUMO gap tends to underestimate the experimental values.

The primary objective of this computational study was to construct statistical models with which polymer electronic properties could be predicted using data extracted from computationally inexpensive DFT calculations. These statistical models could then be used in conjunction with other tools developed in previous works as the beginnings of a multistage screening process for the efficient discovery of new materials for optoelectronic applications.

Computational Methods

Monomer Data Set Selection

Initially, 100 diverse thiophene monomers were selected by choosing a variety of substitutions, including simple functional groups, aromatic and non-aromatic fused ring systems. Bulky substitutions (e.g. t-butyl) were avoided to maintain some consistency in the degree of planarity among the set of conjugated oligothiophene copolymers. The monomers in this study were limited to species containing C, H, N, O, S, Se, F, Br, and Cl. Nine monomers which contained Se were eventually excluded from the data set due to lack of MMFF94 parameters for the conformer generation process. The remaining 91 monomers containing C, H, N, O, S, F, Br, and Cl were used as the basis for all data collection and analysis in this study. The structures of these 91 thiophene monomers are shown in **Figure S1** of the supporting information. The monomer SMILES codes are listed in **Table S1** of the supporting information.

Dimer Data Set Selection

The 4095 dimers in this experiment were chosen by combining each compound in the monomer data set into every possible unique dimer pair where order did not matter and repetitions were not allowed. This combination scheme is illustrated in **Figure 1**.

Tetramer, Hexamer, and Octamer Data Set Selection

All tetramers, hexamers, and octamers in this study were constructed from the dimer data set and are of the form $(AB)_n$, where "$(AB)_n$" is the alternating co-oligomer, "AB" is one of the 4095 dimer combinations, and "n" is the number of dimer repeat units. All 4095 tetramer combinations (n = 2) were included in the study. A random selection of 120 diverse dimer combinations was used to generate hexamer (n = 3) data, and 70 of these 120 combinations were randomly selected to generate octamer (n = 4) data. To ensure the small subset of compounds in the octamer set were representative of the entire data set, the distribution and range of the dimer HOMO/LUMO energies of the 120 combinations used to build hexamers and the 70 combinations used to build octamers were compared to the dimer energies of the entire data set. **Figure S2** of the supporting information contains a scatter plot illustrating this comparison. The dimer, tetramer, hexamer, and octamer energy data for the 70 combinations in the octamer subset were used to extrapolate the infinite polymer energy data for these 70 compounds. Only 60 of these 70 compounds were used to construct the predictive OLS models for the polymer energies. The remaining 10 compounds were used as a validation set.

Generation of Optimized 3D Structures

For each compound in this study, the 3D structure was generated from its SMILES[20] code using Open Babel[21,22] and the MMFF94 forcefield[23-29] (steepest descent, 1000 steps, $1x10^{-4}$ kcal/mol convergence criteria). Next, a weighted-rotor search was performed with Open Babel to obtain a low energy conformer (MMFF94, 250 conformers, 25 geometry optimization steps). The structure was further optimized with conjugate gradients (MMFF94, 500 steps, $1x10^{-6}$ kcal/mol). These pre-optimized structures were converted to Gaussian09[30] input files, and their structures were optimized in the gas phase using DFT with the B3LYP[31,32] functional at the 6-31G* level[33]. The HOMO and LUMO eigenvalues derived from density functional theory have no physical meaning and cannot be directly taken as ionization potentials or electron affinities respectively. However, previous studies have shown that eigenvalues calculated with the B3LYP hybrid functional have a favorable correlation with experimental electron affinities[34,35,36], ionization potentials[34], and band gaps[37].

Statistical Methods

Statistical analyses were performed in python[38,39] using the IPython command shell[40] with the NumPy and SciPy[41], Pandas[42], and Statsmodels[43] packages. Plots were made with the Matplotlib[44] graphics environment.

Results and Discussion

Predicting Dimer HOMO and LUMO energies

Frequently, one may refer to a "donor" or "acceptor" monomer, in regards to an electron donor with high HOMO orbital energy and an electron acceptor with low LUMO orbital energy. Unfortunately, since the HOMO-LUMO gaps of the two monomers may be different, the real picture is more complex. **Figure 2** illustrates four common orbital mixing motifs that repeatedly occur in this data set. In cases (a) and (c), one monomer is both the "HOMO donor" and "LUMO donor", while the other is both the "HOMO acceptor" and "LUMO acceptor". In cases (b) and (d), differences in E_g energies lead to one monomer acting as the "HOMO donor" and "LUMO acceptor" while the other acts as the "HOMO acceptor" and "LUMO donor". Case (a) exemplifies ideal behavior with respect to Hückel molecular orbital (HMO) theory. After combining the orbital mixing behaviors. In case (c), the dimer molecular orbitals exhibit a mixing pattern that is characteristic of neither electron delocalization nor localization. Large dipole moments lead to charge transfer interactions that dominate the observed dimer molecular orbital energies. For example, the calculated dipole moment of the dimer in case (c) is 4.9 Debye, compared to 1.7 Debye for the dimer in case (a). In case (d), the dimer molecular orbitals are localized, despite forming a conjugated aromatic system.

It was expected that a statistical relationship exists between the HOMO, LUMO, and E_g energies of the co-dimers and the monomers which compose them. Operating under this assumption, ordinary least squares (OLS) regressions were used to develop predictive models for the co-dimers' HOMO, LUMO

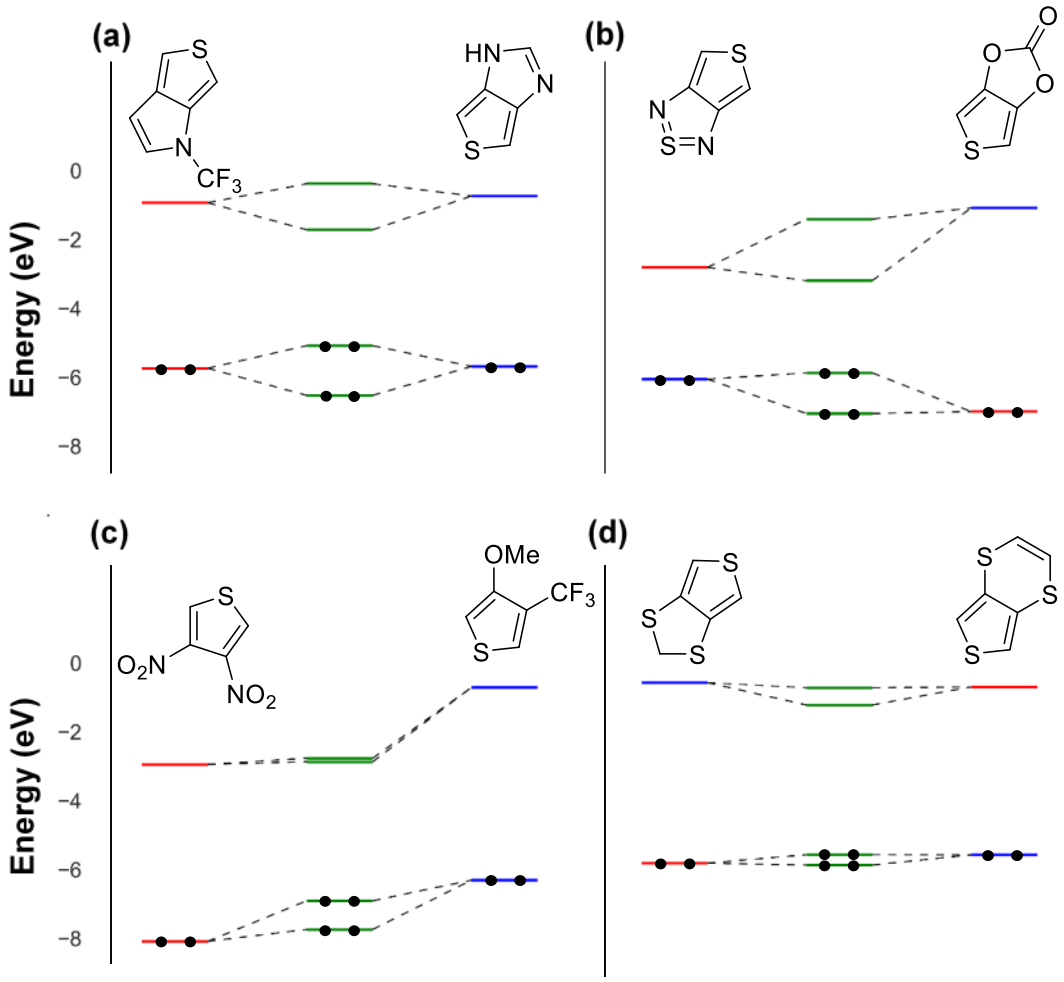

Figure 2. Molecular orbital diagrams of four diverse dimers in the data set. "Donor" orbitals are highlighted in blue. "Acceptor" orbitals are highlighted in red. Dimer orbitals are highlighted in green and filled orbitals are denoted with black circles. Four general orbital mixing motifs are evident in this data set: (a) delocalization, (b) polarization, (c) charge transfer, and (d) localization. Note that differences in monomer gap energies can result in the same monomer being a "HOMO donor" and "LUMO acceptor."

two monomers, their electron densities become delocalized across the conjugated dimer. This leads to an increase in the dimer HOMO energy, a decrease in the dimer LUMO energy and a decrease in the overall E_g energy. In case (b), large difference in monomer E_g energies lead to polar, localized dimer molecular orbitals. Cases (c) and (d) display unexpected (but common) and E_g energies from their respective monomer building blocks. The OLS regression models were constructed using a minimal number of model parameters to maximize the correlation between the dependent and independent variables (R^2). The p-values for the independent variables in each model were minimized (p-value < 0.05) to ensure that the contribution of

each model parameter was statistically significant at the 95% confidence interval. The mean absolute error was calculated for each model by comparing the experimental results from the training set with the predicted values. The models predicting dimer HOMO, LUMO and E_g energies in this experiment are summarized in **Table 1**. The equation and summary of the OLS results for each model are given in **Tables S2 – S7** of the supporting information.

There are obvious correlations between monomer and dimer frontier molecular orbital energies. It should be noted that the when the dimer E_g energy is simply approximated as the difference in energy between the HOMO donor's HOMO energy and the LUMO acceptor's LUMO energy, the errors are relatively large (0.50 +/- 0.33 eV, not shown in **Table 1**). When this difference is treated as a parameter in a univariate OLS model, the errors are drastically reduced (0.22 +/- 0.17 eV). As expected, the multivariate OLS model that contains HOMO, LUMO and E_g data from both monomers further decreases the model error (0.18 +/- 0.14 eV). These statistical models neglect to consider orbital interaction and delocalization, which is captured by even simple quantum models such as HMO theory. A simple Hückel secular determinant could be used to calculate resonance integrals for each monomer pair after dimer formation. It was hypothesized that these resonance integrals could be treated as a parameter to increase the accuracy of the dimer models in **Table 1**.

Hückel Analysis

The extent of orbital interaction in each dimer combination under study was analyzed using simple HMO theory. If each thiophene monomer in every dimer pair is treated as a single delocalized unit, then the relative interaction between each monomer can be calculated with the secular determinant shown in **Equation 1**.

$$\begin{vmatrix} \alpha_A - E & \beta_{AB} \\ \beta_{AB} & \alpha_B - E \end{vmatrix} = 0 \qquad \textbf{(1)}$$

where "α_A" is the HOMO energy of monomer A, "α_B" is the HOMO energy of monomer B, and "β_{AB}" is the resonance integral representing the extent of interaction between the HOMO orbitals of monomers A and B when combined into dimer AB. The solution of this determinant yields two energy values: E_1 is equal to the HOMO energy of dimer AB, and E_2 is equal to the HOMO -1 energy of dimer AB. Solving **Equation 1** for β_{AB} yields **Equation 2**.

$$\beta_{AB} = \frac{1}{2}\sqrt{(E_1 - E_2)^2 - {\alpha_A}^2 - {\alpha_B}^2 + 4\alpha_A\alpha_B} \qquad \textbf{(2)}$$

The dimer and monomer HOMO DFT data collected in this study, in conjunction with **Equation 2,** allowed for the calculation of β_{AB}^{HOMO}. This quantity is a measure of the relative delocalization of electron density between the HOMO orbitals of the monomers in each of the 4095 dimer pairs. Using **Equation 2** and substituting the LUMO energy of monomer A, the LUMO energy of monomer B, the LUMO+1 energy of dimer AB, and the LUMO energy of dimer AB for α_A, α_B, E_1 and E_2 respectively, β_{AB}^{LUMO} was also calculated for each of the 4095 dimer pairs.

The β values calculated in HMO theory are negative valued. Large negative β values imply the monomers interact with high delocalization of electron density across the molecule, while less negative β values ($\beta \rightarrow 0$) indicate that electrons are more localized. A large portion of the β_{AB}^{HOMO} and β_{AB}^{LUMO} values that were calculated from the thiophene data set were complex numbers. This suggests that the simple secular determinant shown in **Equation 1** is poorly described by HMO theory. The diversity of orbital mixing motifs illustrated in **Figure 2**, particularly the case of significant charge transfer interactions, likely contribute to the abundance of complex β values. Histograms of $(\beta_{AB}^{HOMO})^2$ and $(\beta_{AB}^{LUMO})^2$ are shown in **Figure 3**. The distribution of these β integrals can be partly explained by examining differences in the HOMO or LUMO orbital energies of monomer units prior to dimer formation.

Table 1. Summary of the statistical correlation and error associated with each dimer model

Model	R^2	Mean Absolute Error (eV)	Standard Deviation of Error (+/- eV)
$HOMO_{Dimer}$ ~ $HOMO_{HD}$	0.76	0.22	0.16
$HOMO_{Dimer}$ ~ $HOMO_{HD}$ + $HOMO_{HA}$	0.90	0.14	0.11
$LUMO_{Dimer}$ ~ $LUMO_{LA}$	0.88	0.18	0.13
$LUMO_{Dimer}$ ~ $LUMO_{LD}$ + $LUMO_{LA}$	0.94	0.12	0.10
$E_{g\ Dimer}$ ~ ($LUMO_{LA}$ - $HOMO_{HD}$)	0.70	0.22	0.17
$E_{g\ Dimer}$ ~ $E_{g\ HD}$ + $E_{g\ HA}$ + $HOMO_{HD}$ + $LUMO_{LA}$	0.81	0.18	0.14

$_{HD}$ = The monomer with the higher energy HOMO, "HOMO donor"
$_{HA}$ = The monomer with the lower energy HOMO, "HOMO acceptor"
$_{LD}$ = The monomer with the higher energy LUMO, "LUMO donor"
$_{LA}$ = The monomer with the lower energy LUMO, "LUMO acceptor"
E_g = LUMO – HOMO, "HOMO/LUMO Gap"

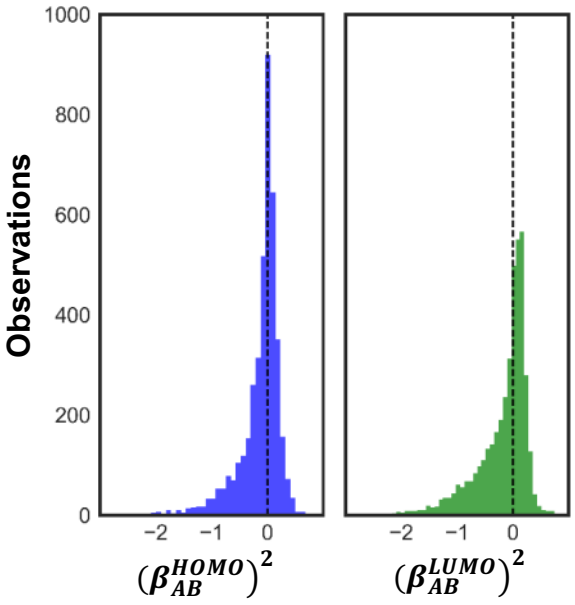

Figure 3. Distribution of calculated β^2 values. Note that for many cases, $\beta^2 < 0$ due to charge transfer effects.

The difference in monomer HOMO energies and the difference in monomer LUMO energies are defined as "Δ_{HOMO}" and "Δ_{LUMO}" respectively. The LUMO en ergy of a co-dimer will experience a greater contribution from the monomer with the lower lying LUMO orbital. Conversely, the dimer HOMO energy will experience a greater contribution from the monomer with the higher lying HOMO orbital. The difference between these molecular orbital energies is defined as the "Dimer shift". An illustration of these definitions is shown in **Figure 4**.

The dimer shifts of the LUMO orbital energies were examined as a function of Δ_{LUMO}. The $(\beta_{AB}^{LUMO})^2$ values were also looked at as a function of Δ_{LUMO}. Analogous relationships were explored for the monomer and dimer HOMO orbitals. These models are shown in **Figure 5**.

Figure 5 (b) and (d) offer some insight into the origins of the complex β values that were generated by **Equation 2**. As Δ_{HOMO} and Δ_{LUMO} increase in magnitude, the secular

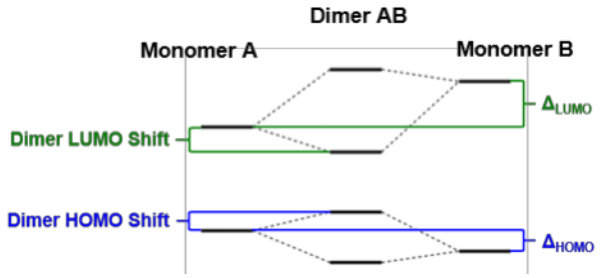

Figure 4. Definitions of "Δ_{HOMO}", "Δ_{LUMO}", "Dimer LUMO shift" and "Dimer HOMO shift". "Δ_{HOMO}" is defined as the difference in monomer HOMO energies. "Δ_{LUMO}" is defined as the difference in monomer LUMO energies. The "dimer HOMO shift" is defined as the higher monomer HOMO energy subtracted from the dimer HOMO energy. The "dimer LUMO shift" is defined as the dimer LUMO energy subtracted from the lower monomer LUMO energy. Note that the dimer HOMO or LUMO shift can be positive or negative depending on the orbital mixing motif. "Δ_{HOMO}" and "Δ_{LUMO}" are always positive.

determinant shown in **Equation 1** does an increasingly poor job at describing the molecular orbital interactions of the monomers after dimer formation. Once the difference in monomer LUMO energies is greater than 1.5 eV, nearly all calculated β_{AB}^{LUMO} values are complex numbers. This is also the case for large differences in monomer HOMO energies ($\Delta_{HOMO} \geq 1.3$ eV).

It is evident in **Figure 5** (a) that when the difference in monomer LUMO energies is large (~1.5-2 eV), the dimer LUMO energies begin to *increase* in energy relative to the LUMO acceptor monomer's LUMO energy. **Figure 5** (c) illustrates that modest differences in monomer HOMO energies (less than 1 eV) frequently result in dimer HOMO energies that *decrease* in energy relative to the HOMO donor monomer's HOMO energy. Increases in LUMO energies and decreases in HOMO energies after extending the conjugated π systems contradict expected FMO energy trends. The negative HOMO shifts were much more prevalent than the negative LUMO shifts, therefore the HOMO shift trends were further explored.

It was hypothesized that large dipole moments lead to charge transfer interactions, which cause the abundance of unexpected negative HOMO shifts illustrated in **Figure 5**. To test this hypothesis, the dataset was partitioned into three subsets: small HOMO shifts (-0.2 to 0.2 eV), positive HOMO shifts (> 0.2 eV), and negative HOMO shifts (< -0.2 eV). Within each subset, the dimer HOMO shift was plotted versus calculated dimer dipole moment using joint kernel density estimate (KDE) plots. The joint KDE plots for the positive HOMO shift and negative HOMO shift subsets are shown in **Figure 6**. While there is little correlation between the dimer HOMO shift and the dimer dipole moment, there are clear differences between the average dipole moments in each subset. The negative HOMO shift compounds have an average dipole moment of 5.6 $\pm$ 2.0 D (725 compounds). The positive HOMO shift compounds have an average dipole moment of 3.3 $\pm$ 1.9 D (1402 compounds). The compounds with small HOMO shifts (not shown in **Figure 6**) have an average dipole moment of 4.1 $\pm$ 2.0 D (1968 compounds). Only the extreme dimer HOMO shift values are shown in **Figure 6** to more clearly establish the relationship with the dimer dipole moments.

The lack of direct correlation between dimer HOMO shifts and dimer dipole moments imply that other factors are involved in this trend, but there is an obvious tendency for dimers with very high dipole moments to exhibit unusual shifts in HOMO energy after monomer combination. This is consistent with charge transfer interactions dominating molecular orbital energies. Incorporation of the $(\beta_{AB}^{HOMO})^2$, $(\beta_{AB}^{LUMO})^2$, Δ_{HOMO}, Δ_{LUMO}, HOMO shift, LUMO shift, or dimer dipole moment parameters into the statistical models of **Table 1** failed to improve their predictive power.

Infinite Polymer Extrapolation

Predicted polymer energy values were calculated by plotting the orbital energies of the oligomers as a function of the inverse of their chain length. A linear OLS best fit line was generated for each of these plots and extrapolated to the y-axis (n = ∞) to

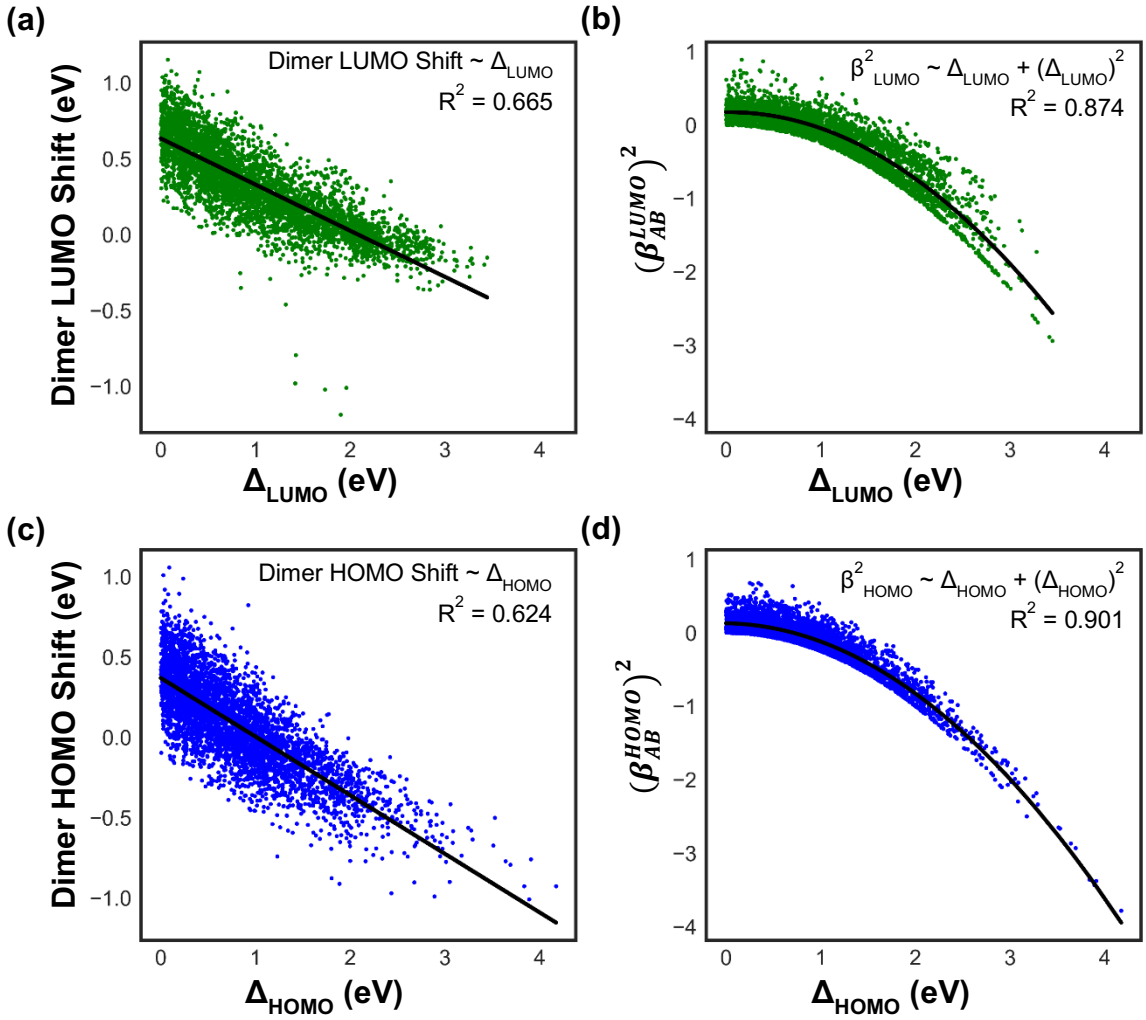

Figure 5. Dimer FMO energy shifts and calculated β^2 values as a function of monomer LUMO (a, b) and HOMO (c, d) energy gaps.

estimate the infinite polymer HOMO, LUMO and E_g energies. When utilizing the simple 1/n extrapolation method to determine the E_g energy of a polymer from its constituent oligomers, the predicted infinite polymer gap energy tends to underestimate the experimental energy due to the saturation effects of the π conjugated system[16,18,19]. These saturation effects were neglected to simplify the analysis. The 1/n model was employed to estimate the polymer HOMO, LUMO, and E_g energies for the 70 compounds in the octamer data set, and these values were used as the basis for the predictive models developed below.

Predicting Polymer HOMO and LUMO energies

Following the logic presented when constructing the models to predict dimer FMO energies, it was expected that similar relationships exists between the HOMO, LUMO, and E_g energies of the $(AB)_n$ co-oligomers, and the building blocks which compose them (i.e. "A" and "B" monomers, "AB" dimers, "ABAB" tetramers, etc.). It was also expected that the relative change in HOMO, LUMO, or E_g energy from dimer to tetramer would be a useful predictor for the extrapolated polymer FMO energies. OLS regression models were constructed by utilizing a minimal number of parameters to maximize the R^2 value while minimizing the p-values of each parameter. These models are summarized in **Table 2**. The tetramer and dimer models were constructed using the data from all 4095 compounds. The polymer models were constructed using only 60 of the 70 possible compounds, the other 10 compounds were used as a validation set. The mean absolute errors in **Table 2** were calculated by comparing the experimental results from the training set with the predicted values. The equation and summary of the OLS results for the tetramer and polymer models are given in **Tables S8 – S13** of the supporting information.

The mean absolute errors and standard deviations of errors given in **Tables 1** and **2** are biased estimates. The training set data was compared to the model predictions in each case. The polymer FMO energies of a validation set containing 10 compounds were compared to the values predicted by the polymer models in **Table 2** to give a more accurate picture of the errors associated with these models. The results of this comparison are given in **Table 3**.

Table 2. Summary of the statistical correlation and error associated with each model

Model	R^2	Mean Absolute Error (eV)	Standard Deviation of Error (+/- eV)
*$HOMO_{Dimer} \sim HOMO_{HD} + HOMO_{HA}$	0.90	0.14	0.11
*$LUMO_{Dimer} \sim LUMO_{LD} + LUMO_{LA}$	0.94	0.12	0.10
*$E_{g\ Dimer} \sim E_{g\ HD} + E_{g\ HA} + HOMO_{HD} + LUMO_{LA}$	0.81	0.18	0.14
$HOMO_{Tetramer} \sim HOMO_{Dimer} + HOMO_{HD}$	0.91	0.14	0.11
$LUMO_{Tetramer} \sim LUMO_{Dimer} + LUMO_{LA}$	0.93	0.13	0.11
$E_{g\ Tetramer} \sim E_{g\ Dimer} + HOMO_{HD} + LUMO_{LA}$	0.70	0.20	0.16
†$HOMO_{Polymer} \sim (2*HOMO_{Tetramer} - HOMO_{Dimer})$	0.91	0.11	0.11
†$LUMO_{Polymer} \sim (2*LUMO_{Tetramer} - LUMO_{Dimer})$	0.95	0.06	0.06
†$E_{g\ Polymer} \sim (2*E_{g\ Tetramer} - E_{g\ Dimer})$	0.91	0.16	0.14

*Models from **Table 1,** included for comparison

†Only 60 of the 70 compounds for which polymer data was obtained were used to construct these models

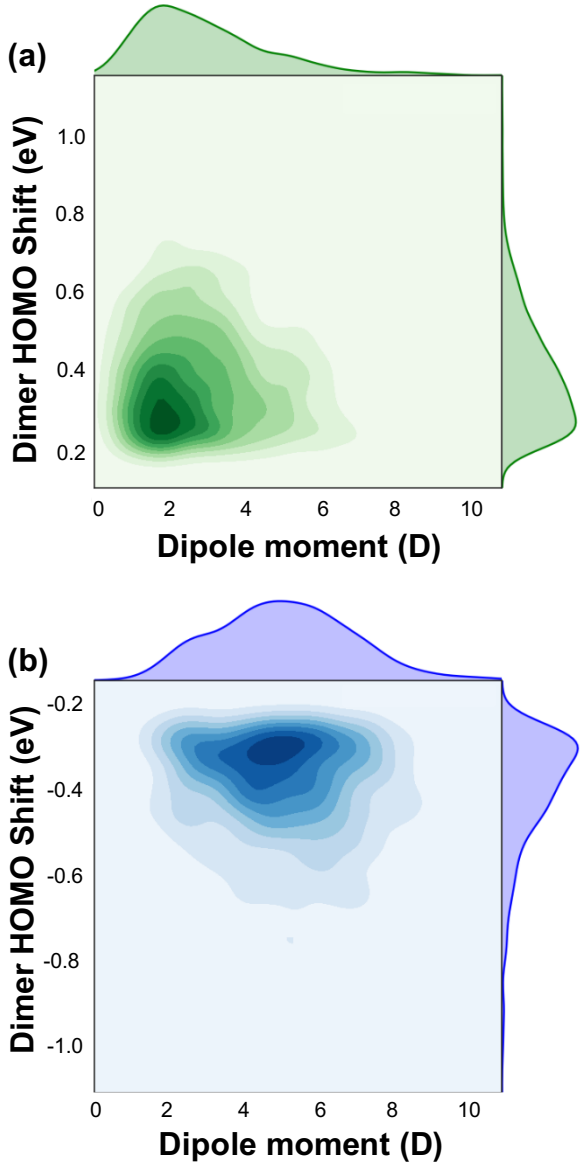

Figure 6. Joint Kernel Density Estimate plots for the subsets of dimers with HOMO shifts greater than 0.2 eV (a) and less than -0.2 eV (b). The average dipole moment for the dimers with positive HOMO shifts greater than 0.2 eV is 3.3 ± 1.9 D. The average dipole moment for the dimers with negative HOMO shifts less than -0.2 eV is 5.6 ± 2.0 D.

Table 3. Validation set error comparisons

Polymer orbital	Data Set	MAE (eV)	STD (+/- eV)
HOMO	Training	0.11	0.11
	Validation	0.21	0.17
LUMO	Training	0.06	0.06
	Validation	0.09	0.09
E_g	Training	0.16	0.14
	Validation	0.26	0.22

The non-biased cross-validation sets have higher errors than the training sets, but these errors are still relatively small (e.g. 0.26 eV MAE for the polymer E_g). This suggests some over-fitting may occur, as expected due to the relatively modest training set size of 60 compounds.

Model Analysis

To determine how well the stepwise application of these statistical models perform, the infinite polymer molecular orbital energies were calculated using the functions generated by the OLS regressions along with 1) experimental monomer DFT energies only, 2) experimental monomer and dimer DFT energies, and 3) experimental dimer and tetramer DFT energies only. The mean absolute error for each case was computed by comparing these results to the polymer FMO energies. This process was performed for the polymer HOMO energy, the polymer LUMO energy, and the polymer E_g energy. The results of this analysis are summarized in **Table 4**. Note that the training set data was used to calculate the predicted polymer FMO energies in each case for ease of comparison. The actual magnitudes of the errors in **Table 4** are likely larger.

It is evident that the stepwise application of these statistical models using only monomer and/or dimer DFT energies results in relatively large errors when calculating the infinite polymer HOMO, LUMO or E_g energy. A drastic reduction in the MAE occurs once the tetramer data is incorporated into the calculation. This analysis suggests that using only monomer energy data to make predictions about the electronic energy of a copolymer would lead to impractically large errors.

Table 4. Statistical model error comparisons

Polymer orbital	Data used to test model	MAE (eV)	STD (+/- eV)
HOMO	Monomer	0.39	0.31
	Monomer and Dimer	0.27	0.21
	Dimer and Tetramer	0.11	0.11
LUMO	Monomer	0.31	0.23
	Monomer and Dimer	0.27	0.22
	Dimer and Tetramer	0.06	0.06
E_g	Monomer	0.56	0.38
	Monomer and Dimer	0.49	0.36
	Dimer and Tetramer	0.16	0.14

Conclusions

This work suggests that the standard donor-acceptor model is a relatively naïve one. The monomer acting as the HOMO donor is not necessarily the LUMO donor, which tends to complicate molecular orbital interactions upon combination. The diversity of molecular orbital interactions after dimer formation in this large dataset evidence the shortcomings of the donor-acceptor model. After dimer formation, some HOMO energy shifts were as high as 0.8 eV, suggesting a large delocalization of electron density. Most of the dimers in this data set exhibited an unexpected negative HOMO energy shift after dimer formation (2458 out of 4095 compounds). This observation is attributed to charge transfer interactions that result from large molecular dipole moments. When the Hückel model was applied to the compounds in this data set, the majority of the calculated β values were *complex* numbers. This is likely a reflection of the fact that the Hückel model does not account for charge transfer interactions. The complex β values became much more prevalent when the differences between monomer HOMO and LUMO orbital energies were large, as evidenced in **Figure 4**. These large differences between monomer HOMO and LUMO orbital energies are also related to negative dimer HOMO and LUMO energy shifts and charge transfer interactions.

The statistical models in this study performed relatively poorly until tetramer data were used to predict polymer FMO energies. Even using the calculated dimer FMO energies from the *training set* to calculate the polymer FMO energies, the mean absolute error for the HOMO/LUMO gap was on the order of 0.5 eV – the actual errors would likely be even higher. The inclusion of tetramer data drastically decreased the mean absolute error of the statistical models (HOMO/LUMO gap error of 0.16 eV for the training set, 0.22 eV for the test set). This error reduction after the inclusion of tetramer data likely follows from the emergence of a more accurate picture of oligomer conformational preferences. The conformation along the backbone of a polymer plays a significant role in the extent to which molecular orbital interactions can occur across conjugated systems. These conformational preferences are difficult, if not impossible, to predict using only monomer and dimer data.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported in part by the University of Pittsburgh Center for Research Computing through the computational resources provided.

Notes and references

1 X. Guo, M. Baumgarten and K. Müllen, *Prog. Polym. Sci.*, 2013, **38**, 1832-1908.
2 L. Bian, E. Zhu, J. Tang, W. Tang and F. Zhang, *Prog. Polym. Sci.*, 2012, **37**, 1292-1331.
3 Y. Liu, T. T. Larsen-Olsen, X. Zhao, B. Andreasen, R. R. Søndergaard, M. Helgesen, K. Norrman, M. Jørgensen, F. C. Krebs and X. Zhan, *Sol. Energy Mater. Sol. Cells*, 2013, **112**, 157-162.
4 N. Hundt, K. Palaniappan, P. Sista, J. W. Murphy, J. Hao, H. Nguyen, E. Stein, M. C. Biewer, B. E. Gnade and M. C. Stefan, *Polym. Chem.*, 2010, **1**, 1624.
5 Z.-G. Zhang and Y. Li, *Sci. China: Chem.*, 2014, **58**, 192-209.
6 J.-F. Lutz, *Polym. Chem.*, 2010, **1**, 55.
7 I. Y. Kanal, J. S. Bechtel and G. R. Hutchison, in *Sequence-Controlled Polymers: Synthesis, Self-Assembly, and Properties*, American Chemical Society, 2014, Chapter 25, pages 379-393.
8 K. Sivula, C. K. Luscombe, B. C. Thompson and J. M. J. Fréchet, *J. Am. Chem. Soc.*, 2006, **128**, 13988-13989.
9 Y. Sun, X. Lu, S. Lin, J. Kettle, S. G. Yeates and A. Song, *Org. Electron.*, 2010, **11**, 351-355.
10 R. A. Janssen and J. Nelson, *Adv. Mater.*, 2013, **25**, 1847-1858.
11 J.-L. Reymond, R. van Deursen, L. C. Blum and L. Ruddigkeit, *Med. Chem. Comm.*, 2010, **1**, 30.
12 I. Y. Kanal, S. G. Owens, J. S. Bechtel and G. R. Hutchison, *J. Phys. Chem. Lett.*, 2013, **4**, 1613-1623.
13 S. S. Zade and M. Bendikov, *Org. Lett.*, 2006, **8**, 5243-5246.
14 S. S. Zade, N. Zamoshchik and M. Bendikov, *Acc. Chem. Res.*, 2011, **44**, 14-24.
15 H. Kuhn, *J. Chem. Phys.*, 1949, **17**, 1198-1212.
16 J. Torras, J. Casanovas and C. Aleman, *J. Phys. Chem. A*, 2012, **116**, 7571-7583.
17 R. E. Larsen, *J. Phys. Chem. C*, 2016, **120**, 9650-9660.
18 H. Meier, U. Stalmach and K. H., *Acta. Polym.*, 1997, **48**, 379-384.
19 J. Gierschner, J. Cornil and H. J. Egelhaaf, *Adv. Mater.*, 2007, **19**, 173-191.
20 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31.
21 N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 5.
22 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
23 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490.
24 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 520.
25 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 553.
26 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 616.

27 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 720.
28 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 730.
29 T. A. Halgren and R. B. Nachbar, *J. Comput. Chem.*, 1996, **17**, 587.
30 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N. J. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian 09 (Revision A.02), Gaussian, Inc., Wallingford, CT, 2009.
31 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648-5652.
32 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785-789.
33 R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724-728.
34 C. G. Zhan, J. A. Nichols and D. A. Dixon, *J. Phys. Chem. A*, 2003, **107**, 4184-4195.
35 J. C. Rienstra-Kiracofe, C. J. Barden, S. T. Brown and H. F. Schaefer III, *J. Phys. Chem. A*, 2001, **105**, 524-528.
36 G. de Oliveira and J. M. L. Martin, *Phys. Rev. A*, 1999, **60**, 1034-1045.
37 J. Muscat, A. Wander and N. M. Harrison, *Chem. Phys. Lett.*, 2001, **342**, 397-401.
38 T. E. Oliphant, *Comput. Sci. Eng.*, 2007, **9**, 10-20.
39 K. J. Millman and M. Aivazis, *Comput. Sci. Eng.*, 2011, **13**, 9-12.
40 F. Perez and B. E. Granger, *Comput. Sci. Eng.*, 2007, **9**, 21-29.
41 S. van der Walt, S. C. Colbert and G. Varoquaux, *Comput. Sci. Eng.*, 2011, **13**, 22-30.
42 W. McKinney, *Proc. of the 9th Python in Science Conf.*, 2010, 51-56.
43 S. Seabold and J. Perktold, *Proc. of the 9th Python in Science Conf.*, 2010, 57-61.
44 J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90-95.