Accelerated Computational Discovery of High-Performance Materials for Organic Photovoltaics by Means of Cheminformatics

Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics

Roberto Olivares-Amaya[a], Carlos Amador-Bedolla[a,b], Johannes Hachmann[a], Sule Atahan-Evrenk[a], Roel S. Sánchez-Carrera[a]*, Leslie Vogt[a] and Alán Aspuru-Guzik[a]†

[a]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St, Cambridge, MA 02138, USA

[b] Facultad de Química, Universidad Nacional Autónoma de México, México, DF 04510, México

† aspuru@chemistry.harvard.edu

Abstract

In this perspective we explore the use of strategies from drug discovery, pattern recognition, and machine learning in the context of computational materials science. We focus our discussion on the development of donor materials for organic photovoltaics by means of a cheminformatics approach. These methods enable the development of models based on molecular descriptors that can be correlated to the important characteristics of the materials. Particularly, we formulate empirical models, parametrized using a training set of donor polymers with available experimental data, for the important current-voltage and efficiency characteristics of candidate molecules. The descriptors are readily computed which allows us to rapidly assess key quantities related to the performance of organic photovoltaics for many candidate molecules. As part of the Harvard Clean Energy Project, we use this approach to quickly obtain an initial ranking of its molecular library with 2.6 million candidate compounds. Our method reveals molecular motifs of particular interest, such as the benzothiadiazole and thienopyrrole moieties, which are present in the most promising set of molecules.

1 Introduction

Current human consumption of energy amounts to 550 EJ per year, which corresponds to 260 million barrels of oil equivalent (MBOE) per day. If the world economy keeps growing at rates close to what has been observed in the last hundred years, human consumption of energy will reach 360 MBOE per day by 2035.[1] To maintain a supply for this growing demand is a challenge, primarily because of the decreasing energy return on investments. At the same time, the continuing use of fossil fuels will increase the impact of global climate change. Almost 87% of the energy consumed by humanity is currently derived from fossil fuels[2] and all renewable energy sources will be needed in order to satisfy the present and future demand for clean energy.

Solar power is a prominent source for renewable energy, in particular for the production of electricity without greenhouse gas emissions. Solar cells are made of thin layers of photovoltaic materials which can harness sunlight for conversion into electricity. Crystalline silicon-based solar cells have dominated the field of commercial photovoltaics, but drawbacks in the manufacturing process as well as high production cost have precluded them from widespread use.[3] Thin-film technologies have led to the development of solar cells based on other inorganic materials such as CdTe,[4] as well as the development of dye-sensitized solar

*Present Address: Robert Bosch LLC Research and Technology Center, 1 Cambridge Center, Cambridge, MA, 02142

cells.[5] Although none of these technologies have reached a higher efficiency than crystalline silicon at 25%,[6] they allow for the possibility of cheaper fabrication and a favorable efficiency/cost ratio, as their production process is less energy-intensive.

Organic photovoltaic (OPV) cells represent another thin-film approach which has drawn a lot of attention and has shown significant progress in recent years.[7] OPVs are particularly promising due to the abundance of their main constituents, their low cost, scalability, and versatility of their installation. Moreover, the potential of rational design to improve the performance of the solar cells has driven recent progress in OPVs. The record power conversion efficiencies of OPVs have improved considerably in the last years: from 1% in 1985;[8] 4% in 2002; 6% in 2009; and up to 9.2% in 2011.[9] If power conversion efficiencies of 10-15% in combination with a lifetime of more than 10 years can be achieved in production materials, OPVs could compete with inorganic-based photovoltaics and become a commercially viable alternative for harnessing electricity from sunlight in a wide range of applications.

Organic-optoelectronic materials span a vast chemical space due to the structural versatility of their carbon-based framework. The prospect of exploring this space has interesting implications for materials design considerations. Due to challenges in the synthesis and experimental characterization of these systems, usually only a modest number of compounds can be studied as candidates for active materials in OPVs.[10,11] Approaches that involve the *in silico* screening of potential organic semiconductors for OPV applications can aid in accelerating the discovery of high-efficiency materials.[12–14]

In this perspective, we review the recent progress of semiconductor polymers for plastic solar cells, and later present the basic ideas of cheminformatics (chemical informatics) for the search of novel organic photovoltaic materials. We adopt the use of physicochemical and topological descriptors, which are commonly known and employed in drug discovery, for the identification of promising organic semiconductors with desired current-voltage characteristics and high power conversion efficiencies. In this context we discuss the systematic construction and optimization of the descriptor models. This technique is employed as part of the Harvard Clean Energy Project,[14,15] a high-throughput *in silico* screening and design effort to develop novel high-performance materials for OPVs. The cheminformatics investigation presented here is a valuable complement to the much more time consuming first-principles electronic structure calculations performed in other parts of this project.

2 Bulk-Heterojunction Solar Cells

The state-of-the-art of OPVs are based on a bulk-heterojunction (BHJ) architectures of two semiconductor compounds: one acting as a electron donor (typically a polymer, or a small molecule) and the other acting as an electron acceptor (a high electron affinity molecule).[16] Fig. 1 shows a schematic illustration of a BHJ solar cell. The photovoltaic process begins with light absorption and ends with charge transport to the electrodes. It occurs through the following steps: i) optical absorption and exciton formation, ii) exciton migration, iii) exciton dissociation at the donor-acceptor interface, iv) charge carrier migration to the electrodes, and v) charge collection at the electrodes. These five steps are summarized in Fig. 1(b). This mechanism naturally carries potential losses at each stage, mainly stemming from inefficient absorption in the beginning and exciton recombination at the intermediate steps. Further details of these elementary processes and their limiting factors have been described extensively in the literature.[17–19]

The parameters which determine the overall efficiency of the energy conversion process in a solar cell are examined in terms of its current-voltage (*i.e.*, power) characteristics.[20,21] The power conversion efficiency (PCE) is defined as the percentage of the ratio of power output (P_{out}), to power input (P_{in}). P_{out} is the maximum (m) obtainable electric power: the product of current, J_m, and voltage, V_m. It is also possible to define P_{out} as depending linearly on the product of the short circuit current density (J_{sc}), the open circuit voltage, (V_{oc}), and the fill factor (FF). The fill factor is the ratio of the maximum power, $J_m V_m$, to the product of J_{sc} and V_{oc}. The product $J_m V_m$ represents the potential power available under the ideal conditions imposed by $J_{\text{sc}} V_{\text{oc}}$.[22] The FF then becomes a parameter that measures the capacity of the device to obtain the most power available. Losses depend on the parasitic resistance of the device and other inefficiencies, which are related to the cell morphology.[23,24] Thus, the formula to compute power conversion

efficiency can be written as:

$$\%\mathrm{PCE} = \frac{\mathrm{FF} \cdot J_{\mathrm{sc}} \cdot V_{\mathrm{oc}}}{P_{\mathrm{in}}} \times 100. \quad (1)$$

J_{sc} and V_{oc} are quantities that can easily be determined under device illumination and largely depend on the molecular properties of the donor and acceptor moieties.

As detailed by Brabec,[21] V_{oc} is related to exciton dissociation, which leads to the charge separation process (step iii, above). V_{oc} scales linearly with respect to the energy difference between the highest occupied molecular orbital (HOMO) of the donor and the lowest unoccupied molecular orbital (LUMO) of the acceptor.[25] J_{sc}, on the other hand, largely depends on the charge mobility and the bandgap of the donor, which determines the spectral overlap: the smaller the bandgap, the higher the spectral overlap. The theoretical understanding of the important parameters for high photovoltaic efficiency has led to models that predict the efficiency of a donor material with respect to a given acceptor, commonly PCBM (1-(3-methoxycarbonyl)propyl-1-phenyl-[6,6]C_{61}), as a function of their energy levels.[26,27] In particular, the model of Scharber *et al.*, Ref. [26], has been instructive for this purpose due to its simplicity.

The first generation of OPV architectures involved a structure in which donor and acceptor layers of O(100nm) were spin-cast. These original designs for donor-acceptor bilayers are limited by the intrinsic exciton diffusion length, as the excitons formed in the donor layer have to reach the interface with the acceptor for the exciton to dissociate.[28] BHJ devices involve blends of donor and acceptor materials which mix at the nanometer scale, creating connected domains of O(10 nm) of donor and acceptor materials that facilitate exciton diffusion to the interface before recombination takes place.[16,29,30] A challenge for theoretical methods for materials discovery is that the ultimate efficiency of BHJ materials depends on annealing conditions and co-solvents, also known as additives.[31] The general complexity and multiscale nature of the device morphology is very hard to model with electronic structure theory.

Recent developments in device architecture that go beyond simple BHJ designs are numerous. They include textured substrates for increased light path lengths,[32] the addition of a titanium oxide (TiO_x) layer on top of the BHJ layer as an optical spacer which has internal quantum efficiencies of 100%[33,34] and other improvements such as plasmonic concentrators.[35]

3 Organic Photovoltaic Materials

In this section, we provide a brief overview on the evolution of the different design approaches for novel OPV materials. The sequence of developments will be relevant to the discussion of our computational approach in the following sections, as the results from the cheminformatics screening should correspond to the experimentally observed trends. We show a (by no means exhaustive) overview of the OPV milestones in Table 1.

Many of the initial donor materials for BHJ devices derived from poly[2-methoxy-5-(3',7'-dimethyloctyloxy)-1,4-phenylene vinylene] (MDMO-PPV, Fig. 2). These donors are combined with PCBM as the acceptor. PCBM has been extensively used as a solar cell acceptor material, along with its C_{70} analogue,[16,36,37] and all reported values in this perspective (*e.g.,* V_{oc}, J_{sc}, the FF and PCE) use these molecules as acceptors. MDMO-PPV has a low-lying HOMO of -5.4 eV. For the junction, a V_{oc} of 0.82 V and a J_{sc} of 5–6 mA/cm^2 was measured. The small J_{sc} value can be explained by the large donor bandgap, and it ultimately limits the PCE to 3.3%.[7,38]

Regioregular poly-(3-hexylthiophene) (rr-P3HT Fig. 2) with a 1.9 eV bandgap emerged as a predominant donor due to its higher J_{sc}, and refined morphological characteristics that lead to a presumably higher exciton mobility than found in MDMO-PPV. This advantage results in efficiencies of over 5%[33,37] A high lying donor HOMO precludes this molecule from having a larger PCE, despite the improvements in J_{sc} and morphology.

Recent searches for donor materials have focused on improving either V_{oc} or J_{sc}, while eq. 1 clearly suggests the need to optimize both. However, there seems to be a trade-off between J_{sc} and V_{oc} that can partially be attributed to the relatively high LUMO of the fullerene-based acceptors and their interaction with the frontier molecular orbitals of the donor. To improve upon this, a new generation of co-monomer

based materials was introduced, in which an electron-donor and an electron-acceptor motif are coupled to form the "monomer" of the polymer unit.

Donor-acceptor designs— One strategy, first proposed by Havinga, Zhang and others.,[39–41] involves improving the donor polymer properties by using a set of alternating electron-rich (*i.e.*, donor) and electron-deficient (*i.e.*, acceptor) moieties to form co-monomers. This approach results in a smaller bandgap for the donor via the hybridization of the energy levels between the donor (typically with high HOMO) and the acceptor (low LUMO) fragments in the co-monomer.[41,42] It also improves the intramolecular charge-transfer.[43–45] This technique is thus labeled "donor-acceptor polymer" approach (DADA in Table 1). For example, Mühlbacher *et al.* synthesized poly-[2,6-(4,4-bis-(2-ethylhexyl)-4H-cyclopenta[2,1-b;3,4-b']- dithiophene)-alt-4,7-(2,1,3-benzothiadiazole)] (PCPDTBT Fig. 2), which shows a low bandgap (1.7 eV) and also absorption activity in the infrared region, with an overall efficiency of 3.2%.[46] Morphological improvements via the co-solvent approach mentioned above, led to an improved efficiency of 5.5%.[31]

Incorporation of high-mobility inducing fragments— Further improvements have been achieved by adding a moiety which promotes the charge-carrier mobility. The co-monomer poly-[2,7-(9-(2'-ethylhexyl)-9-hexyl-fluorene)-alt-5,5-(4',7')-di-2-thienyl-2',1',3'-benzothiadiazole] (PFDTBT Fig. 2)[47,48] has three components: thiophene (T), as a donor; benzothiadiazole (BT), as an acceptor; and fluorene (F), as the high-mobility fragment; it is represented as DTAT in Table 1. The fluorene moiety is known to absorb at short wavelengths, but the mixture with thienyl fragments red-shifts the absorption. The thiophene moiety has in addition good hole-transport properties, increases planarity, and is used as the fragment on which alkyl chains are typically fastened for improved processing. Initial successes with this design was demonstrated by an improved $V_{\rm oc}$ of 1.05 V. However, the $J_{\rm sc}$ remained low at 3.65 mA/cm^2. The PCE of PFDTBT was estimated to be 1.7%, but modifying the R groups in fluorene pushed the PCE to 2.1%.[47,48]

Blouin *et al.* concentrated on finding better acceptor units in the co-monomer. They substituted the fluorene moiety for carbazole, a fully aromatic system, to obtain poly-[N-9'- heptadecanyl-2,7-carbazole-alt-5,5-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT, Fig. 2). A considerable increase of the $J_{\rm sc}$ to 6 mA/cm^2 was achieved, albeit with a slightly lower $V_{\rm oc}$ of 0.9 V. This resulted in an overall PCE of 3.6%.[49] Following a similar technique, Wang *et al.* replaced the fluorene moiety with silafluorene (SiF), and obtained a PCE of 5.4%.[50] The higher efficiency is a result of the broader absorption spectrum of silafluorene, which allows for a $J_{\rm sc}$ of 9.5 mA/cm^2.

Blouin *et al.* continued to optimize their PCDTBT co-monomer to focus on the acceptor fragments.[10] Despite the HOMO and LUMO levels being ideally tuned in some cases, they were unable to improve their reported PCE of 3.6%.[49] Recently, Park *et al.* were able to obtain 6.1% efficiency using PCDTBT and adding a titanium oxide (TiO_x) layer on top of the BHJ layer as an optical spacer;[34] an example of optimization via modifying the device architecture. This improvement brings PCDTBT efficiency close to the PCE limit predicted by the Scharber model.[26]

Chen and Cao reviewed and analyzed donor-acceptor polymer materials which contained either benzothiadiazole, thiophene, thienopyrazine or quinoxaline for a total of 39 co-monomers.[51] Their analysis revealed that systems with a lower bandgap resulted in higher PCE values. The authors argue that although there is a linear trend between the HOMO position and the $V_{\rm oc}$,[25] there is significant scatter in the data to conclude that other effects influence the open-circuit voltage. Furthermore, Yang *et al.*[52] have studied the effect of alkyl chains on the $V_{\rm oc}$ and $J_{\rm sc}$ for a given backbone. They found that there is a significant change with respect to length and shape of the R-groups, which argues in favor of strong dependence on morphology. The relationship between these changes are attributed to the strength of intermolecular interactions between the polymer and PCBM blend.

Weak donor, strong-acceptor motifs— Zhou and Price *et al.* have extended the donor-acceptor polymer approach by introducing the concept of *weak*-donors and *strong*-acceptors.[53–57] Co-monomers are built similar to PFDTBT: donor, thiophene, acceptor, thiophene (DTAT in Table 1). Once again, thiophene moieties are present to increase planarity and as a location to add the R-groups. Zhou *et al.* generate a *weak*-donor moiety by starting from a strong (*i.e.*, electron-rich) component like thiophene and then fusing it with benzene, a less electron-rich moiety.[53] In the case of *strong*-acceptors, it is important for the moiety to be π-electron deficient: the benzene moiety in the benzothiadiazole unit can for instance be substituted

with pyridine, to generate thiadiazolo[3,4-*c*]pyridine.[54] Power conversion efficiencies following this design have reached up to 6.3%. Further work revealed an explicit dependence between the donor HOMO and the acceptor LUMO of this co-monomer layout. Recent findings by these authors have concentrated on optimizing these co-monomers using different acceptors (including fluorinated moieties) and resulted in a PCE of over 7%.[55,57]

Quinoidal structures— A successful technique to reduce the bandgap was based on using alternating thieno[3,4-*b*]thiophene (TTP) and benzodithiophene (BDT) units.[58–61] The reduction of the bandgap is due to the stabilization of the quinoidal structure of TTP. BDT experiences quinoidal stabilization as well, but it also provides rigidity to the backbone (represented as QUINO in Table 1. Liang, Chen *et al.* have also explored the use of alkoxy sidechains to yield further improvements. In particular, poly[4,8-bis(2-ethylhexyloxy)-benzo[1,2-b:4,5-b']dithiophene-2,6-diyl-alt-(4-octanoyl-5-fluoro-thieno[3,4-b]thiophene-2-carboxylate)-2,6-diyl] (PBDTTT-CF, Fig. 2) has the TTP unit alkoxylated and fluorinated in positions 2 and 3, respectively. PBDTTT-CF results in a maximum PCE of 7.7%,[60] which is one of the best efficiencies to date.

The fluorine moiety shifts the donor HOMO and LUMO values which leads to a greater V_{oc}, without affecting the J_{sc}. Further work based on the quinoidal strategy has met with mixed results since the control of the bandgap becomes more difficult when there are no explicit acceptors and donors in the polymer.[62]

Improved acceptor materials— The development of acceptor materials has been dominated by functionalized fullerene derivatives, such as PCBM. PCBM has been extensively employed as a solar-cell acceptor material since it was first reported in 1995.[16,36,37] Although the HOMO and LUMO levels of these systems are not ideal for the known donor polymers, no better candidates have been found.[26] C_{60} and C_{70}-based molecules exhibit a high electron mobility and affinity, which is highly isotropic due to their spherical shape. Functionalizing them with the esther moiety provides for good solubility, as well as a higher LUMO level, which reduces its work function.[63] Due to the relative success of PCBM, the search for better organic photovoltaic materials has primarily become a pursuit for finding ideal donor properties constrained by the energy levels of fullerene-based acceptors.

Recent reviews on acceptor materials include those by Anthony on organic-based non-fullerene molecules[64] and by Xu and Qiao on inorganic-based systems.[65] We refer to the recent work by Gendron and Leclerc for other classes of donor polymers.[66] As outlined in this section, a series of compound-design strategies for donors and acceptors and their rationalization has resulted in a systematic increase of reported efficiencies. In the following section we will explore how computational approaches based on cheminformatics can provide guidance towards innovation and the next generation of OPV materials.

4 Cheminformatics Modelling

The rational design of donor and acceptor molecules [1] for OPVs can be pursued by computational studies of potential candidates. Electronic structure calculations represent a valuable tool to characterize optoelectronic features and processes central to the performance of organic semiconductors.[67,68] Current calculation schemes allow the prediction of electronic properties such as HOMO, LUMO and optical bandgap, as well as other molecular properties that are considered to ultimately be related to the OPV efficiency, such as partial charges, intramolecular interactions, and geometries. These calculations are, however, still time consuming and computationally demanding. The Harvard Clean Energy Project[14,15] (CEP) has been set up for an automated, large scale *in silico* characterization of millions of molecules. Based on computational resources provided by distributed volunteer computing in collaboration with IBM's World Community Grid,[69] CEP is currently performing a systematic screening of millions of candidate molecules using electronic structure theory. Cheminformatics methods allow the "transformation of data into information and information into knowledge"[70] and they are being employed as a complementary approach to the quantum chemical work within CEP.

To date, cheminformatics has primarily been designed to provide a fast way of screening large libraries of potential compounds, mostly for pharmaceutical applications. This discipline has been described as

[1] In the following we will collectively use the term 'molecules' for monomers, oligomers, and actual molecules.

"all the information resources that a scientist needs to optimize the properties of a ligand to become a drug." [71] However, the tools developed in this field, and closely related techniques in machine learning and pattern recognition, can in principle be applied to other materials discovery endeavors. Developments in cheminformatics have been driven by the combination of experimental high-throughput screening (the assay and analysis of more than a million chemical reactions) and with the ability to computationally predict physicochemical parameters (called descriptors). The basic strategy of this approach is to obtain these descriptors for candidate molecules, often obtained from designated candidate libraries, [72–80] to score their fitness with respect to a desired set of properties.

One of the most important methods is the identification of quantitative structure–property relationships (QSPR). [81–86] This technique has focused intensely in the search for molecules to be experimentally screened as potential drugs, or as drug leads. [87] More recently, QSPR were employed in the study of certain molecules for an understanding of the fundamental processes of cellular and organismic biology. [88,89] In similar fashion to QSPR, quantitative structure-activity relationships (QSAR) are used to study the biological activity of such problems. We note that the complexities faced in the interactions between organic molecules in biological systems are greater than in those found in organic electronic materials. Despite these challenges, cheminformatics has been successful in several areas on the interface between chemistry and biology. [90] For instance, it is possible to analyze the conformation of drug candidates to evaluate their docking potential to a particular biomolecular target and for a prediction of its use as a pharmacophore.

QSPR have been developed for a wide variety of applications, which include single-molecule, intermolecular and reactive properties. The success of this approach has stimulated its use in recent years, as can be seen in several reviews. [87,91] The materials science community has just begun to utilize machine-learning methods, which encompass cheminformatics and QSPR. Work in this area has led to the prediction of crystal structures of inorganic molecules, [92–95] as well as the development of methods for visualizing and identifying potential porous materials. [96,97]

The simplest QSPR approaches are based on linear regression models, but more sophisticated forms which incorporate genetic algorithms, artificial neural networks, and the Gaussian processes technique have been developed in recent years. [12,98–100] Several other techniques in cheminformatics have been used for the identification of leads not related to regression models. These include statistical tools used in machine learning such as principal component analysis, linear discriminants, and decision trees. [81,84,101,102]

QSAR and QSPR largely rely on the calculation of molecular properties called descriptors, which we will discuss in Sec. 4.1. Descriptors include physical, chemical and topological properties. Descriptors can be classified as either one-, two- or three-dimensional, depending on whether they describe bulk properties, connectivities or conformation-dependent properties, respectively. [83,103] The use of descriptors in cheminformatics has provided simple rules to evaluate druglikeness, as in the case of Lipinski's Rule of Five which analyzes molecules using a set of structural descriptors: molecular mass, hydrogen bond donors, hydrogen bond acceptors, partition coefficient, and number of rotatable bonds. [104] These rules have been very useful for the development of drug leads. Therefore, there have been active efforts to develop computer codes that allow the rapid evaluation of hundreds of such descriptors. [105–107]

4.1 The Molecular Library and Physicochemical Molecular Descriptors

In order to search for donor molecules that have the best combination of electronic properties, we built a molecular library of approximately 2.6 million conjugated molecules. The molecular library employed is built via a combinatorial molecule generation scheme starting from a set of 30 molecular building blocks (in the Supporting Information, SI). The fragments include the most prevalent molecular motifs used in the experimental design of OPVs to date and are chosen with input from experimentalist collaborators from the group of Zhenan Bao at Stanford University to ensure synthetic feasibility. [108] As discussed in Sec. 3, R-groups play an important role in OPV materials but for the present work we chose to focus only on the molecular backbone.

We enumerate the library using a virtual reaction-based approach by either *linking* or *fusing* the fragments together, as shown in Scheme 1. We also extend the size of the co-monomers by properly adding molecular

handles, so they can be *further* linked or fused. Complete details of the molecular library generation will be presented in a separate publication.[109]

We use the previously introduced descriptors for an initial characterization of our molecular library. We employ the Marvin code by ChemAxon.[105] ChemAxon provides a set of over 200 descriptors that are relevant for drug design applications, but they nonetheless proved useful in the application for OPV donor materials. We selected descriptors corresponding to elemental analysis, charge, geometry, and electronic states based on Hückel theory for the study of monomers for use as donor in OPVs. For atomic-based properties, we assessed the maximum, minimum and average value in the molecule. There are a total of 33 descriptors in our model, their classes are listed in Table 2. These can be easily computed for the whole library within a few days on a single workstation.

A specific example of a descriptor that displayed statistically significant correlation is the electrophilic localization energy, L(+), which is an atom-centered property based on the Hückel method: the simplest semiempirical approach for obtaining quantum-mechanical properties of conjugated molecules.[110] L(+) is the energy related to removing an atom from conjugation, effectively donating two π-electrons to the electrophile. The lower the value of L(+), the more reactive the compound. Therefore, a small value of electrophilic localization energy means that the atom contributes little to the overall conjugation of the molecule. The effect is shown in Scheme 2.

4.2 Descriptor Model and Training Set Results

We have chosen a simple linear regression model for this initial investigation. The descriptors chosen above are assembled accordingly and the resulting model is parametrized using a training set of organic monomers with experimentally known current-voltage characteristics. We selected a set of 50 training molecules compiled from the literature.[46,50,51,111–121] These molecules include aliphatic side chains used to control packing structures. The current work is concerned with donor materials of BHJ design, but this method can naturally be applied to other device architectures and materials given the appropriate training set.

As mentioned in Sec. 1, we focus on the four most relevant parameters for the performance characteristics of a solar cell. These are PCE, and its components as expressed in eq. 1: the FF, V_{oc} and J_{sc}. Note that V_{oc} and J_{sc} largely depend on properties intrinsic to the donor and acceptor. FF broadly depends on the morphology and the specific device architectures. We can therefore expect that the molecular descriptors used and the experimental values will show a better correlation for the first two than for the latter. The expression to determine PCE includes all three parameters and its correlation should thus fit in between the others.

The multiple linear regression for the descriptor models with respect to these four parameters was performed using the R code.[122] The correlation, as obtained by the use of the 33 descriptors, varied from very good ($R^2_{V_{\mathrm{oc}}} = 0.96$, $R^2_{J_{\mathrm{sc}}} = 0.92$) or good ($R^2_{\mathrm{PCE}} = 0.89$) to poor ($R^2_{\mathrm{FF}} = 0.66$). We performed a significance test on the descriptors and eliminated the least significant ones which only slightly reduced the precision of the fit (shown in the SI). The significance of the descriptors was obtained from a two-sided t-statistics test. The p-value for each descriptor ranged from 10^{-3} to 10^{-1}.

In order to mitigate the difficulty of predicting the FF from a purely cheminformatics approach, we also built a model for the product $V_{\mathrm{oc}}J_{\mathrm{sc}}$, which is proportional to PCE but only contains parameters well represented in our cheminformatics approach. We summarize the results related to the coefficients of determination (R^2) of the fitting in Table 3. We also present the results of the predicted properties against the measured ones in Fig. 3. As stressed above, it is not unexpected that the parameters which depend on the material properties, V_{oc} and J_{sc}, result in a much better fit than the FF.

The fit resulted in families of significant descriptors that were different for each of the experimental parameters. The best description included 20 descriptors for V_{oc} and $V_{\mathrm{oc}}J_{\mathrm{sc}}$, 18 for J_{sc} and 15 for PCE. Four descriptors are present in the models of each four parameters. We group estimates of these descriptors in Table 4. We notice that each descriptor in this subset has either a positive or a negative correlation for all four values. The separation between estimates is never larger than two orders of magnitude. Therefore, these descriptors form a tight set of estimates that affect each of the parameters in a similar fashion.

As in most machine learning approaches, it is a complicated task to uniquely specify the role of all of individual descriptors for a specific property. Ultimately the combination of all the descriptors in the model is what makes the fit have a relatively good R^2 value of the fits.

4.3 Predictions from Cheminformatics

We now apply the models created in the previous section to the 2.6 million molecules of the candidate library and summarize our findings. The histograms of the obtained results are shown in Fig. 4. In the cases of V_{oc} and J_{sc} (and therefore in $V_{oc}J_{sc}$) there are a considerable number of molecules with predicted values well above the largest observed to date. These molecules constitute the most promising candidates for BHJ donor OPV materials within the presented cheminformatics approach. Some molecules are predicted to have an unrealistic negative value. The fraction of molecules in this situation is small for all parameters except for the FF, which can easily be explained by its relatively poor model. Being mindful of the limitations of the extrapolation, we find that for V_{oc} nearly half of the molecules are predicted to have a value higher than the best of the experimental molecules (1.04 V), and only 0.8% have a negative value; for J_{sc}, 41.5% of the molecules have a value higher than the best experimental, and 8.3% have a negative value; only 1.5% of the molecules have a predicted value of PCE higher than the highest experimental, and the highest value is 10.4%, but there are 43.4% of molecules with a value of the product $V_{oc}J_{sc}$ higher than the highest experimental; these molecules, combined with an appropriate value of the FF (which is not predicted well by these descriptors) could have values of PCE above current records. These results are summarized in Table 5.

We further investigate which of the highest rated molecules have the best value for each of the three current-voltage parameters considered (V_{oc}, J_{sc}, $V_{oc}J_{sc}$). We test if a promising molecule for V_{oc} is also a good candidate for J_{sc} and $V_{oc}J_{sc}$. We selected the top 10% from each group and compared them. We find that molecules predicted to have a high value of V_{oc} only rarely have also a high value of J_{sc}, and *vice versa*. Fig. 5 shows the position in the predicted V_{oc} *vs.* J_{sc} space of the top 10% of molecules from each group: V_{oc}, J_{sc}, and $V_{oc}J_{sc}$. We observe that molecules predicted to have the highest values of the product $V_{oc}J_{sc}$ have mostly a high value of J_{sc} and an average value of V_{oc}, *i.e.* they have a higher overlap with the top values of J_{sc}. This suggests that the search for high efficiency monomoers, is particularly promising with molecules based on motifs present in both the J_{sc} and $V_{oc}J_{sc}$ optimization.

For a more detailed analysis of the results we focus on the top thousand molecules (all following quantities are taken w.r.t. to the top 1000) with the best current-voltage characteristics. For V_{oc}, we notice that these have at least one silicon atom and are built mostly by both linking and fusing the 30 basic fragments. A typical molecule from this set is shown in Fig. 6a. For J_{sc}, silicon atoms are not as common (161 molecules have at least one) but instead selenium-containing heteroatoms are more frequent (313 molecules have at least one) and the thienopyrrole motif is present in 822 molecules. The molecules of this set have a predominantly linked rather than fused backbone. Fig. 6b shows a typical molecule from this set.

Again, the best expected co-monomers for application in heterojunction OPVs correspond, according to this QSPR analysis, to the ones with the highest value of $V_{oc}J_{sc}$, for which the set of the best thousand have molecules with silicon atoms (375), selenium atoms (131), silicon and selenium atoms (53) and come mostly from linking the basic units (890). The benzothiadiazole or pyridinethiadiazole motifs are prevalent in this set of candidates (see Fig. 7), present in 463 molecules. Similarly, units that can potentially have quinoid stabilization are prominent in this set. Specifically, 117 present the thienothiophene moiety. This suggests that the search for monomers with high efficiency as OPV's, should start with molecules based in the motifs presented in Fig. 7, as well with those with potential quinoidal stabilization. We currently work on a cross-validation of the present predictions with the ones from the quantum chemical studies within the Harvard Clean Energy Project which will be presented in the near future.

5 Discussion and Conclusions

In the present work we introduced a cheminformatics based approach for the discovery of promising OPV donor materials. We calculated the current-voltage properties of 2.6 million molecular motifs using linear

regression descriptor models. These allowed us to identify candidates with a favorable set of performance related parameters, which – according to our QSPR analysis – have the prospect of being suitable as high-efficiency BHJ solar cell materials. The molecules with the most promising predictions feature a variety of structural designs, but three motifs appear repeatedly in our top candidates: benzothiadiazole, pyridinethiadazole and thienopyrrole, shown in Fig. 7.

As summarized in Table 1, the evolution of OPV donor materials has followed different design strategies, including the donor-acceptor polymer approach and the quinoidal stabilization. The last column of Table 1 shows the number of molecules, from our top 1000 selection, which are part of each design "generation". We note that PPV-like molecules were not included in our study, and P3HT was not predicted to be in the top 1000. However, there were 17 molecules that followed the donor-acceptor approach. We obtained 13 molecules with the design specified as DTAT, although in our case these systems only had one thiophene scaffold. Finally, molecules with quinoidal stabilization (*i.e.,* containing thienothiophene) numbered 117. A significant fraction of our top molecules hence belongs to the latest generation of OPVs.

Despite the limitations of this simple approach, we can conclude that the use of QSPR and cheminformatics-type approaches can be a valuable guide for the design of lead molecules for solar cell materials. Current efforts to improve upon the presented work include the use of more sophisticated and flexible models, extended and improved training sets, as well as a new generation of descriptors specifically designed for organic semiconductors. The latter can be derived from higher level quantum-chemical studies carried out in the Harvard Clean Energy Project.[15]

6 Acknowledgments

The authors wish to thank Dr. Rajib Mondal and Dr. Anatoliy N. Sokolov for fruitful discussions. We wish to thank IBM for organizing World Community Grid and WCG members for their computing time donations. The CEP is supported by the National Science Foundation through grants No. DMR-0820484 and DMR-0934480 as well as the Global Climate and Energy Project (Stanford grant No. 25591130-45282-A. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of Stanford University, the Sponsors of the Global Climate and Energy Project, or others involved with the Global Climate and Energy Project). R.O.-A was supported by a Giorgio Ruffolo Fellowship in the Sustainability Science Program at Harvard University's Center for International Development, C.A.-B acknowledges financial support from CONACyT México and computer access to DGTIC UNAM, R.S.S.-C thanks the Mary Fieser Postdoctoral Fellowship at Harvard University, and L.V. was supported by an NSF Graduate Research Fellowship. Molecular library generation in this paper were performed on the Odyssey cluster supported by the FAS Research Computing Group. We acknowledge software support from Molecular Networks GmbH, ChemAxon Ltd and Optibrium Ltd.

References

[1] J. J. Conti, P. D. Holtberg, J. A. Beamon, A. M. Schaal, G. E. Sweetnam and A. S. Kydes, *Annual Energy Outlook 2010*, National Energy Information Center, U. S. Energy Information Administration, 2010.

[2] British Petroleum, *BP Statistical Review of World Energy*, `http://bp.com/statisticalreview`, 2011.

[3] R. Gaudiana, *J. Phys. Chem. Lett.*, 2010, **1**, 1288–1289.

[4] A. Slaoui and R. T. Collins, *MRS Bull.*, 2007, **32**, 211–218.

[5] B. O'Regan and M. Grätzel, *Nature*, 1991, **353**, 737–740.

[6] J. Zhao, A. Wang, M. A. Green and F. Ferrazza, *Appl. Phys. Lett.*, 1998, **73**, 1991–1993.

[7] A. J. Heeger, *Chem. Soc. Rev.*, 2010, **39**, 2354–2371.

[8] C. W. Tang, *Appl. Phys. Lett.*, 1986, **48**, 183–185.

[9] R. F. Service, *Science*, 2011, **332**, 293.

[10] N. Blouin, A. Michaud, D. Gendron, S. Wakim, E. Blair, R. Neagu-Plesu, M. Belletête, G. Durocher, Y. Tao and M. Leclerc, *J. Am. Chem. Soc.*, 2008, **130**, 732–742.

[11] R. S. Sánchez-Carrera, S. Atahan, J. Schrier and A. Aspuru-Guzik, *J. Phys. Chem. C*, 2010, **114**, 2334–2340.

[12] N. M. O.Boyle, C. M. Campbell and G. R. Hutchison, *J. Phys. Chem. C*, 2011, **115**, 16200–16210.

[13] A. N. Sokolov, S. Atahan-Evrenk, R. Mondal, H. B. Akkerman, R. S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S. C. Mannsfeld, A. P. Zoombelt, Z. Bao and A. Aspuru-Guzik, *Nat. Commun.*, 2011, **2**, 437–444.

[14] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, A. Gold-Parker, R. S. Sánchez-Carrera, L. Vogt, A. M. Brockway, A. Aspuru-Guzik and the World Community Grid, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.

[15] `http://cleanenergy.harvard.edu`.

[16] G. Yu, J. Gao, J. Hummelen, F. Wudl and A. Heeger, *Science*, 1995, **270**, 1789–1791.

[17] J.-L. Bredas, J. E. Norton, J. Cornil and V. Coropceanu, *Acc. Chem. Res.*, 2009, **42**, 1691–1699.

[18] B. Kippelen and J.-L. Brédas, *Energy Environ. Sci.*, 2009, **2**, 251–261.

[19] H. Borchert, *Energy Environ. Sci.*, 2010, **3**, 1682–1694.

[20] B. C. Thompson and J. M. J. Frechet, *Angew. Chem. Int. Ed.*, 2008, **47**, 58–77.

[21] *Organic Photovoltaics*, ed. C. Brabec, V. Dyakonov and U. Scherf, Wiley-VCH, 2008.

[22] *Organic Photovoltaics: Mechanisms, Materials, and Devices*, ed. S. Sun and N. Sariciftci, CRC, Boca Raton, 2005.

[23] J. Nelson, *The Physics of Solar Cells (Properties of Semiconductor Materials)*, Imperial College Press, 2003.

[24] M.-S. Kim, B.-G. Kim and J. Kim, *ACS Appl. Mater. Interfaces*, 2009, **1**, 1264–1269.

[25] C. Brabec, A. Cravino, D. Meissner, N. Sariciftci, T. Fromherz, M. Rispens, L. Sanchez and J. Hummelen, *Adv. Funct. Mater.*, 2001, **11**, 374–380.

[26] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger and C. J. Brabec, *Adv. Mater.*, 2006, **18**, 789–794.

[27] L. J. a. Koster, V. D. Mihailetchi and P. W. M. Blom, *Appl. Phys. Lett.*, 2006, **88**, 093511.

[28] K. M. Coakley and M. D. McGehee, *Chem. Mater.*, 2004, **16**, 4533–4542.

[29] G. Yu and A. J. Heeger, *J. Appl. Phys.*, 1995, **78**, 4510–4515.

[30] J. Halls, C. Walsh, N. Greenham, E. Marseglia, R. Friend, S. Moratti and A. Holmes, *Nature*, 1995, **376**, 498–500.

[31] J. Peet, J. Y. Kim, N. E. Coates, W. L. Ma, D. Moses, a. J. Heeger and G. C. Bazan, *Nat. Mater.*, 2007, **6**, 497–500.

[32] K. Nalwa, J. Park, K. Ho and S. Chaudhary, *Adv. Mater.*, 2011, **23**, 112–116.

[33] J. Kim, S. Kim, H.-H. Lee, K. Lee, W. Ma, X. Gong and A. J. Heeger, *Adv. Mater.*, 2006, **18**, 572–576.

[34] S. Park, A. Roy, S. Beaupré, S. Cho, N. Coates, J. Moon, D. Moses, M. Leclerc, K. Lee and A. Heeger, *Nat. Photonics*, 2009, **3**, 297–302.

[35] H. A. Atwater and A. Polman, *Nat. Mater.*, 2010, **9**, 205–213.

[36] J. Hummelen, B. Knight, F. LePeq, F. Wudl, J. Yao and C. Wilkins, *J. Org. Chem.*, 1995, **60**, 532–538.

[37] G. Dennler, M. C. Scharber and C. J. Brabec, *Adv. Mater.*, 2009, **21**, 1323–1338.

[38] C. J. Brabec, S. E. Shaheen, C. Winder, N. S. Sariciftci and P. Denk, *Appl. Phys. Lett.*, 2002, **80**, 1288–1290.

[39] E. Havinga, W. Hoeve and H. Wynberg, *Polym. Bull.*, 1992, **29**, 119–126.

[40] Q. Zhang and M. James, *J. Am. Chem. Soc.*, 1998, **120**, 5355–5362.

[41] H. Van Mullekom, J. Vekemans, E. Havinga and E. Meijer, *Mater. Sci. Eng. R*, 2001, **32**, 1–40.

[42] G. Brocks and A. Tol, *J. Phys. Chem.*, 1996, **100**, 1838–1846.

[43] J. Roncali, *Chemical Reviews*, 1997, **97**, 173–206.

[44] S. A. Jenekhe, L. Lu and M. M. Alam, *Macromolecules*, 2001, **34**, 7315–7324.

[45] J. Roncali, *Macromolecular Rapid Communications*, 2007, **28**, 1761–1775.

[46] D. Mühlbacher, M. Scharber, M. Morana, Z. Zhu, D. Waller, R. Gaudiana and C. Brabec, *Adv. Mater.*, 2006, **18**, 2884–2889.

[47] M. Svensson, F. Zhang, S. Veenstra, W. Verhees, J. Hummelen, J. Kroon, O. Inganäs and M. Andersson, *Adv. Mater.*, 2003, **15**, 988–991.

[48] O. Inganäs, M. Svensson, A. Gadisa, F. Zhang, N. Persson, X. Wang and M. Andersson, *Appl. Phys. A*, 2004, **79**, 31–35.

[49] N. Blouin, A. Michaud and M. Leclerc, *Adv. Mater.*, 2007, **19**, 2295–2300.

[50] E. Wang, L. Wang, L. Lan, C. Luo, W. Zhuang, J. Peng and Y. Cao, *Appl. Phys. Lett.*, 2008, **92**, 033307.

[51] J. Chen and Y. Cao, *Acc. Chem. Res.*, 2009, **42**, 1709–1718.

[52] L. Yang, H. Zhou and W. You, *J. Phys. Chem. C*, 2010, **114**, 16793–16800.

[53] H. Zhou, L. Yang, S. Stoneking and W. You, *ACS Appl. Mater. Interfaces*, 2010, **2**, 1377–1383.

[54] H. Zhou, L. Yang, S. C. Price, K. J. Knight and W. You, *Angew. Chem. Int. Ed.*, 2010, **49**, 7992–7995.

[55] H. Zhou, L. Yang, A. C. Stuart, S. C. Price, S. Liu and W. You, *Angew. Chem. Int. Ed. Engl.*, 2011, **50**, 2995–2998.

[56] S. C. Price, A. C. Stuart and W. You, *Macromolecules*, 2010, **43**, 4609–4612.

[57] S. C. Price, A. C. Stuart, L. Yang, H. Zhou and W. You, *J. Am. Chem. Soc.*, 2011, **133**, 4625–4631.

[58] Y. Liang, D. Feng, Y. Wu, S.-T. Tsai, G. Li, C. Ray and L. Yu, *J. Am. Chem. Soc.*, 2009, **131**, 7792–7799.

[59] Y. Liang, Y. Wu, D. Feng, S.-T. Tsai, H.-J. Son, G. Li and L. Yu, *J. Am. Chem. Soc.*, 2009, **131**, 56–57.

[60] H. Chen, J. Hou, S. Zhang, Y. Liang, G. Yang, Y. Yang, L. Yu, Y. Wu and G. Li, *Nat. Photonics*, 2009, **3**, 649–653.

[61] Y. Liang, D. Feng, J. Guo, J. M. Szarko, C. Ray, L. X. Chen and L. Yu, *Macromolecules*, 2009, **42**, 1091–1098.

[62] N. Kleinhenz, L. Yang, H. Zhou, S. C. Price and W. You, *Macromolecules*, 2011, **44**, 872–877.

[63] F. B. Kooistra, J. Knol, F. Kastenberg, L. M. Popescu, W. J. H. Verhees, J. M. Kroon and J. C. Hummelen, *Org. Lett.*, 2007, **9**, 551–554.

[64] J. Anthony, *Chem. Mater.*, 2011, **23**, 583–590.

[65] T. Xu and Q. Qiao, *Energy Environ. Sci.*, 2011, **4**, 2700–2720.

[66] D. Gendron and M. Leclerc, *Energy Environ. Sci.*, 2011, **4**, 1225–1237.

[67] G. R. Hutchison, M. a. Ratner and T. J. Marks, *J. Phys. Chem. A*, 2002, **106**, 10596–10605.

[68] R. S. Sánchez-Carrera, M. C. R. Delgado, C. C. Ferrón, R. M. Osuna, V. Hernández, J. T. L. Navarrete and A. Aspuru-Guzik, *Org. Electron.*, 2010, **11**, 1701–1712.

[69] `http://www.worldcommunitygrid.org`.

[70] F. Brown, *Curr. Opin. Drug Discov. Devel.*, 2005, **8**, 298–302.

[71] F. K. Brown, *Ann. Rep. Med. Chem.*, 1998, **33**, 375–384.

[72] X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *J. Chem. Inf. Comput. Sci*, 1998, **38**, 511–522.

[73] V. J. Gillet, *Methods Mol. Biol.*, 2004, **275**, 335–354.

[74] D. Schnur, B. R. Beno, A. Good and A. Tebben, *Methods Mol. Biol.*, 2004, **275**, 355–378.

[75] U. Fechner and G. Schneider, *J. Chem. Inf. Model.*, 2006, **46**, 699–707.

[76] M. Wang, X. Hu, D. N. Beratan and W. Yang, *J. Am. Chem. Soc.*, 2006, **128**, 3228–3232.

[77] S. Keinan, M. J. Therien, D. N. Beratan and W. Yang, *J. Phys. Chem. A*, 2008, **112**, 12203–12207.

[78] P. S. Kutchukian, D. Lou and E. I. Shakhnovich, *J. Chem. Inf. Model.*, 2009, **49**, 1630–1642.

[79] P. S. Kutchukian and E. I. Shakhnovich, *Expert Opin. Drug Discovery*, 2010, **5**, 789–812.

[80] G. Guntas, C. Purbeck and B. Kuhlman, *Proc. Natl. Acad. Sci.*, 2010, **107**, 19296–19301.

[81] J. Bajorath, *J. Chem. Inf. Comput. Sci*, 2001, **41**, 233–245.

[82] C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **178**, 1616–1626.

[83] L. Xue and J. Bajorath, *Comb. Chem. High Throughput Screening*, 2000, **3**, 363–372.

[84] A. R. Leach and V. Gillet, *An Introduction to Chemoinformatics*, Springer, 2003.

[85] W. L. Chen, *J. Chem. Inf. Model.*, 2006, **46**, 2230–2255.

[86] J. Gasteiger, *Anal. Bioanal. Chem.*, 2006, **384**, 57–64.

[87] B. O. Villoutreix, R. Eudes and M. A. Miteva, *Comb. Chem. High Throughput Screening*, 2009, **12**, 1000–1016.

[88] C. Lipinski and A. Hopkins, *Nature*, 2004, **432**, 855–861.

[89] C. M. Dobson, *Nature*, 2004, **432**, 824–828.

[90] D. Agrafiotis, D. Bandyopadhyay, J. Wegner and H. Van Vlijmen, *J. Chem. Inf. Model.*, 2007, **47**, 1279–1293.

[91] A. Katritzky, M. Kuanar, S. Slavov, C. Hall, M. Karelson, I. Kahn and D. Dobchev, *Chem. Rev.*, 2010, **110**, 5714–5789.

[92] C. C. Fischer, K. J. Tibbetts, D. Morgan and G. Ceder, *Nat. Mater.*, 2006, **5**, 641–646.

[93] K. Rajan, *Annu. Rev. Mater. Sci.*, 2008, **38**, 299–322.

[94] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[95] P. V. Balachandran, S. R. Broderick and K. Rajan, *Proc. R. Soc. A*, 2011, **467**, 2271–2290.

[96] M. Haranczyk and J. Sethian, *J. Chem. Theory Comput.*, 2010, **6**, 3472–3480.

[97] K. Theisen, B. Smit and M. Haranczyk, *J. Chem. Inf. Model.*, 2010, **50**, 461–469.

[98] L. Terfloth and J. Gasteiger, *Drug Discovery Today*, 2001, **6**, 102–108.

[99] F. R. Burden, *J. Chem. Inf. Comput. Sci*, 2001, **41**, 830–835.

[100] O. Obrezanova, G. Csanyi, J. M. R. Gola and M. D. Segall, *J. Chem. Inf. Model.*, 2007, **47**, 1847–1857.

[101] W. Dunn III, D. Scott and W. Glen, *Tetrahedron Comput. Methodol.*, 1989, **2**, 349–376.

[102] C. Kingsford and S. L. Salzberg, *Nature Biotechnol.*, 2008, **26**, 1011–1013.

[103] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, 2002.

[104] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Del. Rev.*, 2001, **46**, 3–26.

[105] ChemAxon, `http://www.chemaxon.com`, 2011.

[106] Stardrop, `http://www.optibrium.com/stardrop`, 2011.

[107] DRAGON, `http://www.talete.mi.it/index.htm`, 2011.

[108] R. Mondal and A. Sokolow, *Personal Communication*.

[109] Olivares-Amaya, R. *et al.*, *In Preparation*, 2011.

[110] N. Isaacs, *Physical Organic Chemistry (2nd Edition)*, Prentice Hall, 1996.

[111] M. Reyes-Reyes, K. Kim and D. Carroll, *Appl. Phys. Lett.*, 2005, **87**, 083506.

[112] L. H. Slooff, S. C. Veenstra, J. M. Kroon, D. J. D. Moet, J. Sweelssen and M. M. Koetse, *Appl. Phys. Lett.*, 2007, **90**, 143506.

[113] R. Mondal, N. Miyaki, H. A. Becerril, J. E. Norton, J. Parmer, A. C. Mayer, M. L. Tang, J.-L. Bredas, M. D. McGehee and Z. Bao, *Chem. Mater.*, 2009, **21**, 3618–3628.

[114] H. Meng, F. Sun, M. Goldfinger, G. Jaycox, Z. Li, W. Marshall and G. Blackman, *J. Am. Chem. Soc.*, 2005, **127**, 2406–2407.

[115] S. Ando, J. Nishida, H. Tada, Y. Inoue, S. Tokito and Y. Yamashita, *J. Am. Chem. Soc.*, 2005, **127**, 5336–5337.

[116] S. Ando, R. Murakami, J. Nishida, H. Tada, Y. Inoue, S. Tokito and Y. Yamashita, *J. Am. Chem. Soc.*, 2005, **127**, 14996–14997.

[117] T. Okamoto and Z. Bao, *J. Am. Chem. Soc.*, 2007, **129**, 10308–10309.

[118] N. Blouin, A. Michaud, D. Gendron, S. Wakim, E. Blair, R. Neagu-Plesu, M. Belletete, G. Durocher, Y. Tao and M. Leclerc, *J. Am. Chem. Soc.*, 2008, **130**, 732–742.

[119] H. Tian, J. Wang, J. Shi, D. Yan, L. Wang, Y. Geng and F. Wang, *J. Mater. Chem.*, 2005, **15**, 3026–3033.

[120] M. Mamada, J. Nishida, D. Kumaki, S. Tokito and Y. Yamashita, *J. Mater. Chem.*, 2008, **18**, 3442–3447.

[121] H. Ebata, E. Miyazaki, T. Yamamoto and K. Takimiya, *Org. Lett.*, 2007, **9**, 4499–4502.

[122] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[123] G. Klopman, J.-Y. Li, S. Wang and M. Dimayuga, *J. Chem. Inf. Comput. Sci*, 1994, **34**, 752–781.

[124] P. T. van Duijnen and M. Swart, *J. Phys. Chem. A*, 1998, **102**, 2399–2407.

[125] K. J. Miller and J. Savchik, *J. Am. Chem. Soc.*, 1979, **101**, 7206–7213.

[126] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar and R. K. Robins, *J. Chem. Inf. Comput. Sci*, 1989, **29**, 163–172.

[127] B. G. Ramsey, *J. Am. Chem. Soc.*, 1965, **87**, 2502–2503.

[128] S. L. Dixon and P. C. Jurs, *J. Comput. Chem.*, 1992, **13**, 492–504.

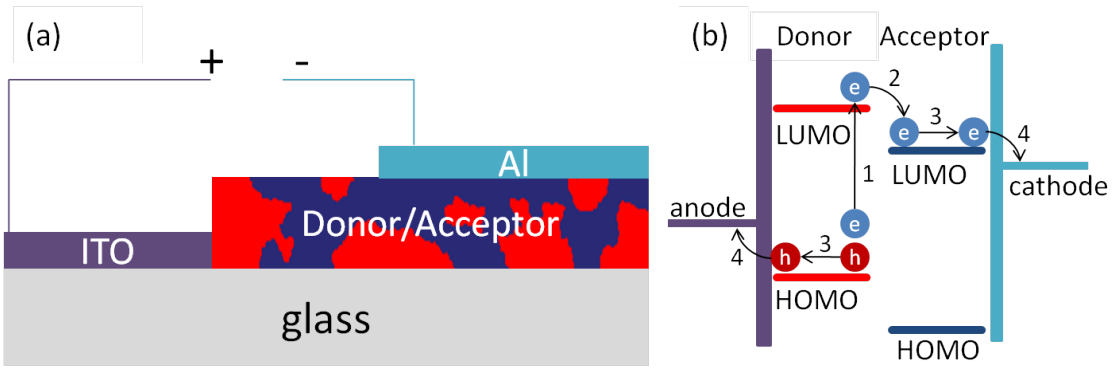


Fig. 1 (a) Device architecture of a bulk-heterojunction solar cell. Light is incident upon the glass substrate. (b) Bulk-heterojunction photophysics: 1) a photon excites an electron to form an exciton, which migrates to the donor/acceptor interface; 2) a difference between the LUMO levels of the donor and acceptor (typically of the order of 300 meV or greater) causes the exciton to dissociate; 3) electrons and holes are transported towards the cathode and anode, respectively; 4) charge is collected at the electrodes, thus transforming light into current.

Table 1 A non exhaustive overview of OPV development. Successive generations of OPV based on monomers and co-monomers: We show record PCE achieved and the number of molecules predicted in the present study corresponding to a particular design concept. Fragment Key: circle, PPV; pentagon, P3HT; square, donor; triangle, acceptor; cross, quinoline. Color Key: Color schematically indicates bandgap size; except gray, which indicates a fragment with good hole mobility.

Generation	Motif	Name	Maximum PCE (experimental)	Top 1000 $V_{oc}J_{sc}$ (predicted)
PPV		MDMO-PPV[38]	3.3%	—
P3HT		P3HT[33]	5.0%	0
DADA		PCPDTBT[31]	5.2%	17
DTAT		PBnDT-DTffBT[55]	7.2%	13
QUINO		PBDTTT-CF[60]	7.7%	117

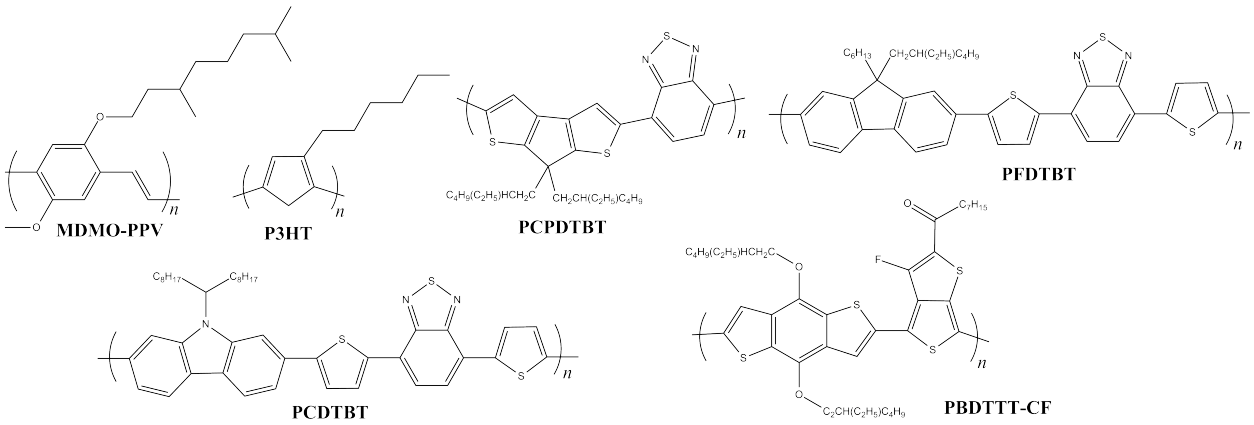
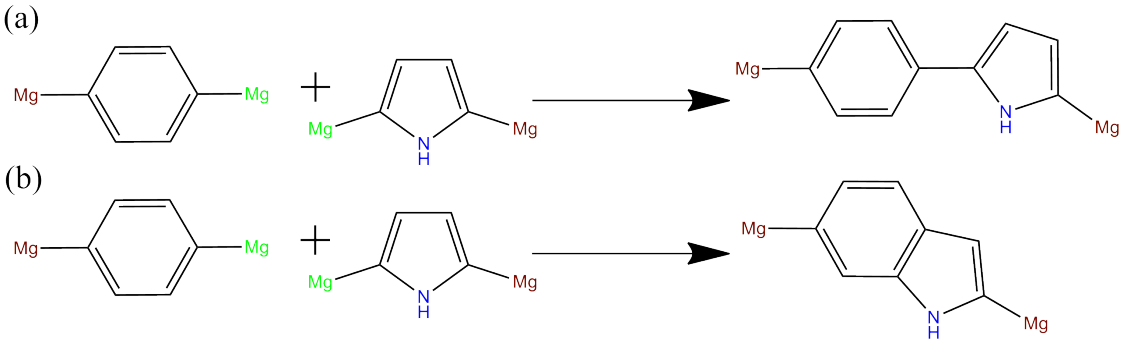


Fig. 2 Chemical structures highlighted in Sec. 3. MDMO-PPV was one of the first donor materials used. P3HT is a prevalent donor which has shown higher J_{sc}, and refined morphological characteristics. PFDTBT and PCDTBT are co-monomers, in which the donor-acceptor polymer strategy was applied to obtain higher PCE. PBDTTT-CF is a co-monomer, which features quinoidal stabilization and with the aid of the fluorine group has yielded an efficiency of 7.7%.



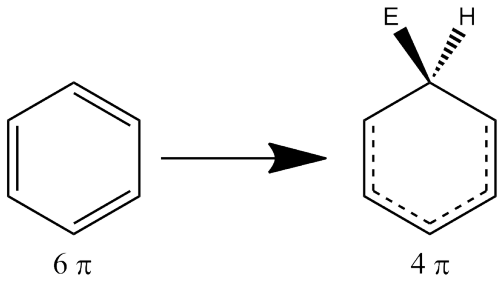
Scheme 1 A reaction-based approach for enumerating a molecular library. (a) Linking reaction: Benzene molecule with Mg chemical handles in the *para* position reacts with pyrrole with Mg chemical handles at the 2,5-position. One set of Mg (green) react to form a linked co-monomer of these moieties. (b) Fusion reaction: Benzene molecule with Mg chemical handles in the *para* position reacts with pyrrole with Mg chemical handles at the 2,5-position to form benzopyrrole. In both cases, a second set of Mg handles (red) is present so that this product can be used as a reagent and enable the generation of co-monomers of greater size.

Table 2 Classes of physicochemical and topological descriptors employed in the presented models. We note that these 17 descriptor classes amount to 33 individual descriptors. An asterisk denotes the descriptor is based on semiempirical Hückel model calculations.

Descriptor	Description
Molecular mass	Molecular mass
logP	Octanol-water partition coefficient, a measure of hydrophobicity based on group contributions from a set of basic fragments fitted to experimental values[123]
Ring count	Number of rings in the molecule
Hydrogen bond acceptor count	Number of hydrogen bond acceptor atoms
Hydrogen bond donor count	Number of hydrogen bond donor atoms
Rotatable bond count	Excludes bonds connecting hydrogens and terminal atoms
Molecular polarizability	Empirical calculation based on a dipole interaction model from atomic polarizabilities, experimental and *ab initio* values[124,125]
Refractivity	Empirical calculation of atomic refractivity; related to London dispersion forces[126]
van der Waals surface area	Molecular surface area as defined by van der Waals radii
van der Waals volume	Molecular surface volume as defined by van der Waals radii
Water accessible area	Water accessible surface area based on atomic properties
Electronic localization energy*	Energy related to removing an atom from conjugation[110,127]
Partial charge*	Partial atomic charges for π systems and electronegativity-based calculation for the σ network[128]
Electron density*	Based on occupancy of atomic-centered orbitals[110,127]
Steric hindrance	Steric hindrance of an atom calculated from the covalent radii values
σ orbital electronegativity*	Mulliken atomic orbital electronegativity from σ orbitals[128]
π orbital electronegativity*	Mulliken atomic orbital electronegativity from π orbitals[128]

Table 3 Summary of linear fitting results for each of the properties we study. We compare the coefficients of determination (R^2) using all 33 descriptors (all desc.) and the statistically significant ones. The number of significant descriptors ranges from 15–20, but the R^2 is not largely affected in all cases.

Property	R^2 (all desc.)	Descriptors	R^2
V_{oc}	0.9580	20	0.9455
J_{sc}	0.9202	18	0.8989
%PCE	0.8937	15	0.8409
FF	0.6567	20	0.6170
$V_{oc}J_{sc}$	0.9025	20	0.8809



Scheme 2 The electrophilic localization energy, L(+): The energy related to the bond formation at a conjugated center, which will remove that center from conjugation, effectively taking away two π-electrons from the conjugated backbone. The lower the value of L(+), the more reactive the compound, meaning that the atom contributes little to conjugation.

Table 4 Estimates for the four prevalent descriptors for the V_{oc} (20 descriptors), J_{sc} (18 descriptors), PCE (15 descriptors), and the product $V_{oc}J_{sc}$ (20 descriptors). The estimate for each of these descriptors are all within two orders of magnitude and have the same sign.

	Estimates			
Descriptor	V_{oc}	J_{sc}	%PCE	$V_{oc}J_{sc}$
Rotatable bond count	+0.2375	+2.3886	+0.8393	+1.9484
Electron density (lowest)	−0.8403	−24.1885	−11.3297	−20.9617
Orbital electronegativity (σ) (average)	−1.4448	−38.4895	−15.5656	−23.4284
Orbital electronegativity (π) (highest)	+0.2317	+2.5199	+1.7823	+3.0837

Table 5 Best current-voltage characteristics predicted from molecular descriptors in the molecular library as compared with experimentally measured for the training set. The highest efficiency predicted is 95% above than the best experimental value. The percentage of molecules predicted to have parameters exceeding the highest experimental value is above 37% for the better fits. Also shown is the percentage of molecules showing (unrealistic) negative values of the parameters.

	V_{oc} (V)	J_{sc} (mA/cm^2)	$V_{oc}J_{sc}$ (mAV/cm^2)	%PCE
Max. value (experimental)	1.04	15.0	8.63	5.32
Max. value (predicted)	2.97	41.5	23.61	10.36
% molecules above highest experimental	43.6	37.2	43.4	1.5
% molecules with negative value (predicted)	0.8	8.3	8.0	19.7

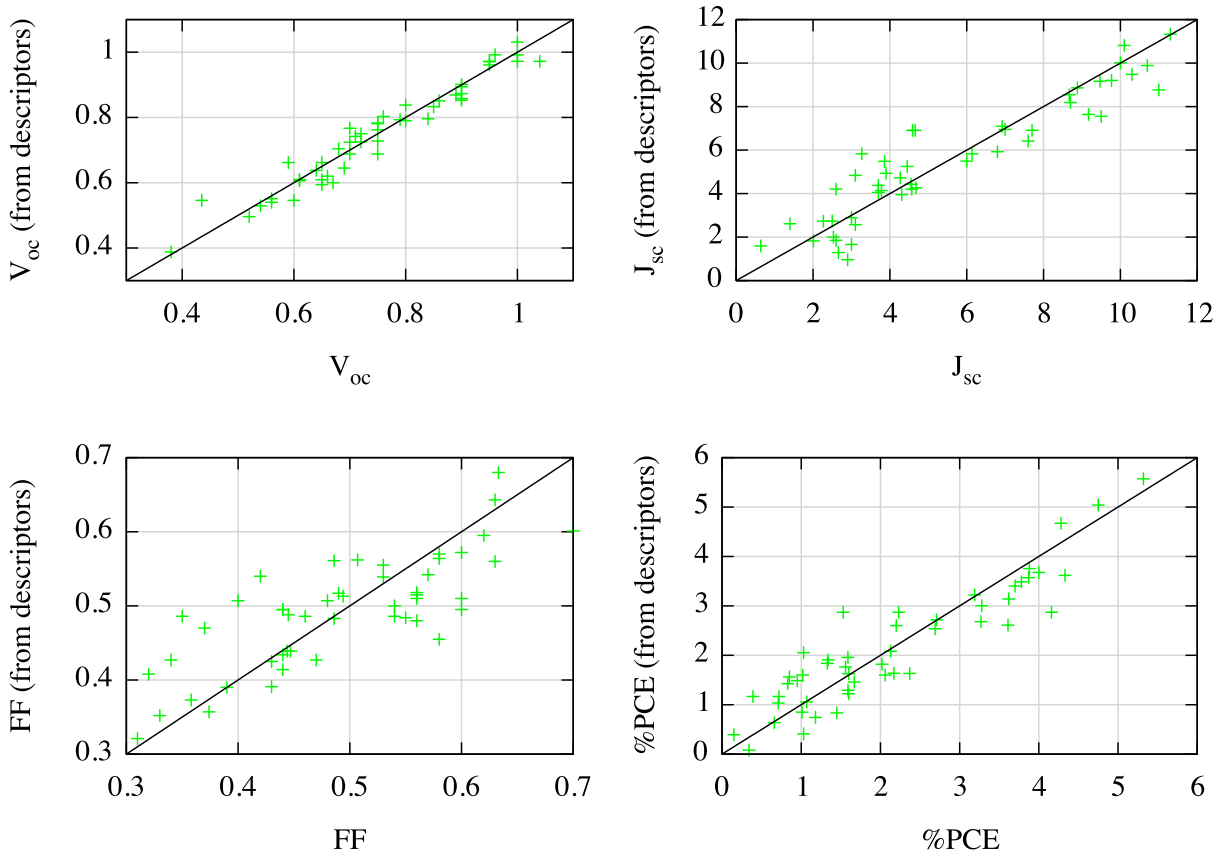


Fig. 3 Results for the multiple linear regression of the four models and the values of the training set. The predicted value from fitted descriptors is compared to the experimental value originally used for fitting. Units are mA/cm^2 for J_{sc}, and V for V_{oc}.

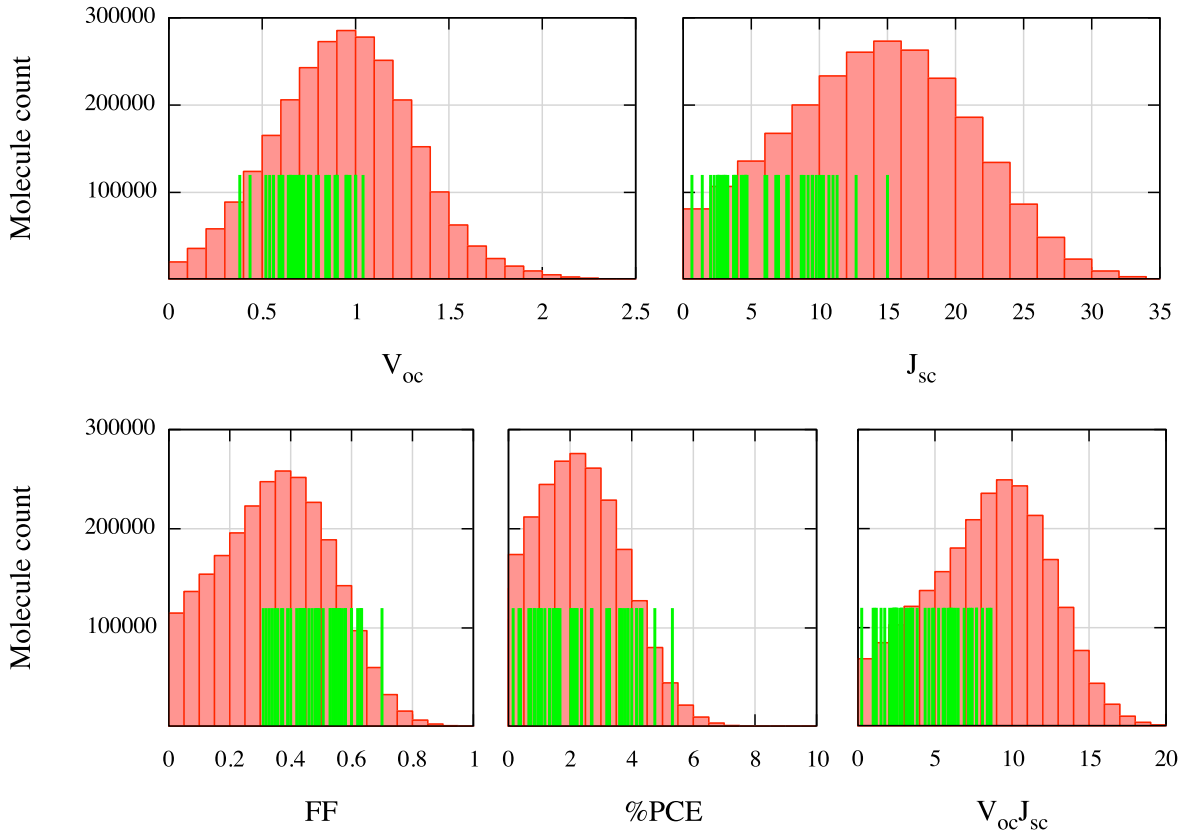


Fig. 4 Histograms of the predicted current-voltage parameters (open circuit voltage ($V_{\rm oc}$), short circuit current density ($J_{\rm sc}$), fill factor (FF), power conversion efficiency (PCE), and the product $V_{\rm oc}J_{\rm sc}$) for the screening of 2,671,405 molecules. Units are mA/cm^2 for $J_{\rm sc}$, and V for $V_{\rm oc}$. The vertical lines correspond to the experimental values of the molecules in the training set (*i.e.*, independent of the y-axis value). Note that the predicted values are larger than the best experimental ones, especially for $V_{\rm oc}$ and $J_{\rm sc}$, their product, and PCE.

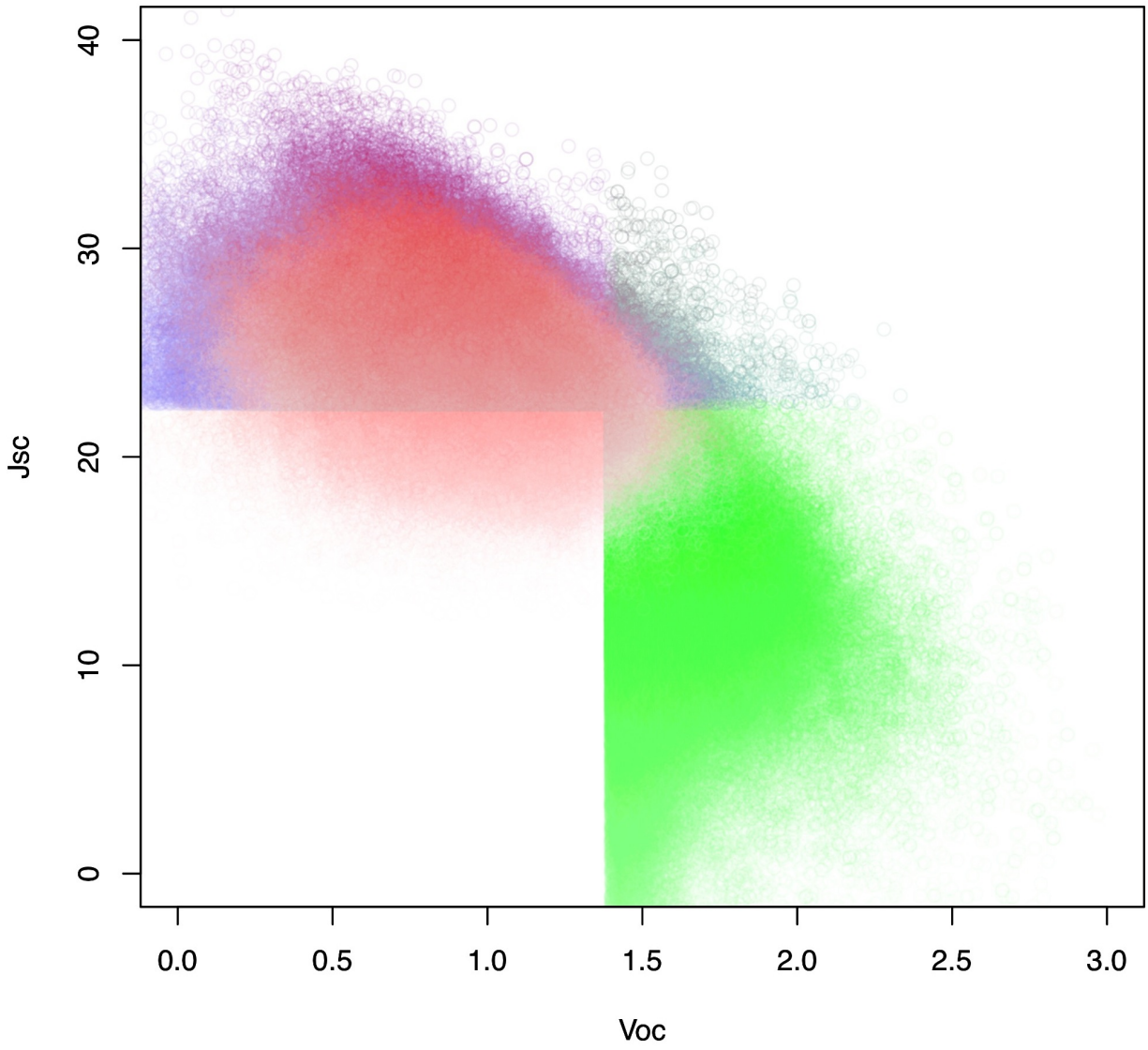


Fig. 5 Top 10% molecules with highest predicted values of V_{oc} (green), J_{sc} (blue), and $V_{\mathrm{oc}}J_{\mathrm{sc}}$ (red). The intensity of the point corresponding to a given molecule is coded according to the value of the product $V_{\mathrm{oc}}J_{\mathrm{sc}}$. The best molecules, according to the present study, are located in the upper left region of the figure. Units are mA/cm^2 for J_{sc}, and V for V_{oc}.

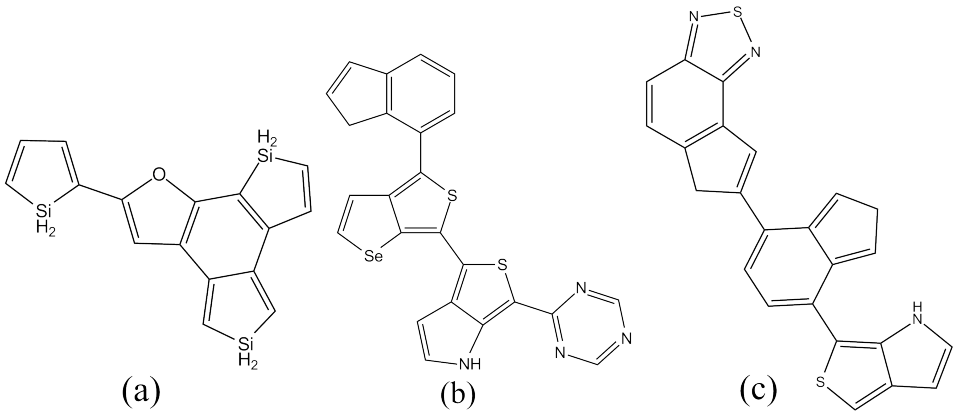


Fig. 6 Typical molecules from the set of cheminformatics predictions for highest (a) V_{oc} (note the mixed linked and fused heterocyclic units with silicon), (b) J_{sc} (note the linked backbone, the selenium atoms and the thienopyrrole motif), (c) $V_{oc}J_{sc}$ (note the mixed linked and fused structure and the benzothiadiazole and thienopyrrole motifs).

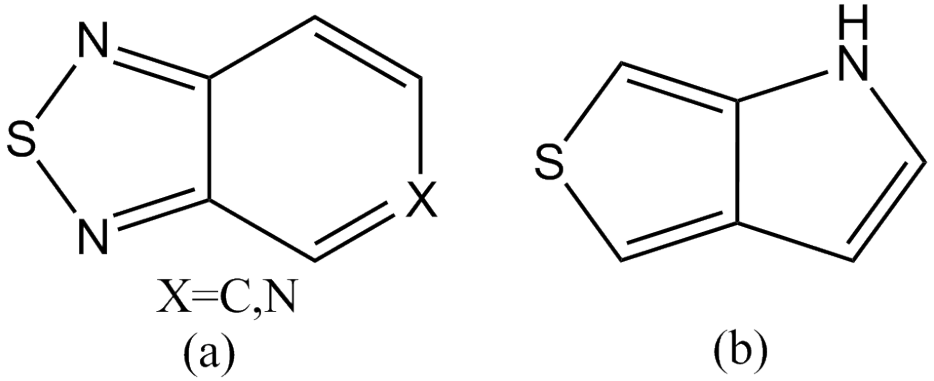


Fig. 7 Ubiquitous motifs present in many of the most promising molecules according to the predicted $V_{oc}J_{sc}$ parameter: a) benzothiadiazole or pyridinethiadiazole motif, b) thienopyrrole motif