Energy & Environmental Science

RSC Publishing

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c3ee42756k

Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project†

Johannes Hachmann,*[ab] Roberto Olivares-Amaya,[ac] Adrian Jinich,[a] Anthony L. Appleton,[d] Martin A. Blood-Forsythe,[a] László R. Seress,[a] Carolina Román-Salgado,[a] Kai Trepte,[a] Sule Atahan-Evrenk,[a] Süleyman Er,[a] Supriya Shrestha,[a] Rajib Mondal,[d] Anatoliy Sokolov,[d] Zhenan Bao[d] and Alán Aspuru-Guzik*[a]

Received 14th August 2013
Accepted 4th October 2013

DOI: 10.1039/c3ee42756k

www.rsc.org/ees

The virtual high-throughput screening framework of the Harvard Clean Energy Project allows for the computational assessment of candidate structures for organic electronic materials – in particular photovoltaics – at an unprecedented scale. We report the most promising compounds that have emerged after studying 2.3 million molecular motifs by means of 150 million density functional theory calculations. Our top candidates are analyzed with respect to their structural makeup in order to identify important building blocks and extract design rules for efficient materials. An online database of the results is made available to the community.

I. Introduction

Organic solar cells are a promising technology for the inexpensive and versatile utilization of solar energy.[1,2] The traditional development of new organic photovoltaic (OPV) materials is predominantly based on empirical intuition or experience with certain compound families. A new design idea is typically followed by a labor-intensive synthesis, characterization, and prototype device optimization. The obtained results are used as feedback for the re-design and improvement of the original candidate. This approach can result in an extended iterative cycle, which may or may not lead to a useful material in the end. Only a small number of structures can thus be tested; the chemical space explored is therefore severely limited, and progress tends to be slow.

These limitations, costs, and the high possibility of failure lead to the idea of a virtual high-throughput prescreening of potential candidate compounds. This approach is devised to facilitate an accelerated development process as efforts can be focused on promising leads while unpromising ones can be excluded early on. Furthermore, the survey of uncharted domains of the molecular space may reveal compound classes with novel and unexpected properties. This notion is inspired by the discovery of fullerenes and carbon nanotubes, and the transformative impact they have had on materials science.[3]

In a recent paper,[4] we introduced the Harvard Clean Energy Project (CEP), an automated, high-throughput framework for the large-scale *in silico* study of molecular materials. It is designed and implemented to identify lead compounds, in particular organic semiconductors for photovoltaic applications. The CEP framework utilizes *first-principles* electronic structure theory (augmented by techniques from chem-informatics/materials informatics and machine learning[5,6]) to characterize millions of molecular motifs and assess their potential. The massive amount of computing time required for this research is provided by distributed volunteer computing by means of IBM's World Community Grid (WCG).[7,8] We note that Hutchison and co-workers have carried out a similar computational screening,[9,10] and that other studies of OPV candidates have been reported in the literature.[11,12] Large-scale initiatives such as the ones by Ceder, Curtarolo, Jacobsen, Nørskov, and Zunger have been successful in exploring the space of inorganic solid state materials.[13–20]

In Section II A we provide an overview of our screening procedure, the data processing, and the current status of the project. Section II B describes our assessment of the candidates by means of the Scharber model and the resulting ranking. In Section II C we discuss the highest-ranked candidates from an empirical perspective, and in Section II D we identify prevalent structural patterns based on a statistical analysis. We

[a]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St, Cambridge, MA 02138, USA. E-mail: hachmann@buffalo.edu; aspuru@chemistry.harvard.edu

[b]Department of Chemical and Biological Engineering, New York State Center of Excellence in Materials Informatics, University at Buffalo, The State University of New York, 612 Furnas Hall, Buffalo, NY 14260, USA

[c]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

[d]Department of Chemical Engineering, Stanford University, 381 North-South Mall, Stanford, CA 94305, USA

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3ee42756k

investigate the correlation between these patterns and the predicted candidate performance in more detail in Section II E. Our findings are summarized in Section III.

II. Results and discussion

A. Virtual high-throughput screening

We have so far examined 2.3 million molecular motifs[21] in 150 million density functional theory (DFT) calculations and generated a data volume of more than 400 terabytes. The CEP represents, to our knowledge, the most extensive quantum chemical investigation ever conducted, and this paper is concerned with the initial analysis of the results to date. The screening is ongoing and we complete about 20 000 workunits every day, each carrying out up to 15 DFT calculations on a candidate.

The 2.3 million compounds characterized up to this point are part of our primary candidate library for OPV donor materials.[4] It was combinatorially generated from 26 basic building blocks according to predetermined rules regarding possible connections. The fragments and connectivity rules were chosen with promise and synthetic feasibility in mind. They are inspired by established moieties from the literature but also include modified ones. Further details about the basic building blocks are given in the analysis section below. We point out that despite these design choices, not all resulting candidates will indeed be accessible for synthesis.

For each of the candidates we generate up to five low-energy conformers using molecular mechanics. These are subsequently subject to geometry optimizations at the DFT level. By considering multiple conformers we attempt to capture possible deviations from the optimum geometry due to interactions in a polymer or bulk material. As detailed in ref. 4, we perform a series of DFT single point calculations on the different geometries using an array of model chemistries (*i.e.*, functional and basis set combinations). The results are empirically calibrated to correct for some of the systematic errors in each theoretical model and account for the situation in a real material. The calibration is based on linear regressions between a training set of known experimental data and the corresponding computational results, *i.e.*, it aligns the predictions of the latter to the former.[4] Details are given in the ESI.† In order to obtain more robust values and reduce random errors introduced by failures of individual model chemistries or calibrations for particular data points, we average over all the independently acquired results. We also average over the different geometries for each candidate. The overall averages can include up to 75 values. The calibrated and averaged results are our best estimates for the quantities of interest.

B. Analysis and ranking *via* the Scharber model

In the following analysis we employ the Scharber model,[22,23] a specialized version of the Shockley–Queisser model for OPVs.[24] The only inputs it requires are the energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). We previously proposed the use of calibrated Kohn–Sham eigenvalues[25–27] in this context and this approach has since been adopted by other groups as well.[11] The statistical analysis of our DFT post-processing is summarized in Table 1, and it reveals a small spread in the values.[28]

We emphasize that the predictions from the Scharber analysis are subject to the limitations of this relatively simple model, its various assumptions, and the quality of the input data provided by the approach described above. The resulting power conversion efficiency (PCE) values should be interpreted as the *potential performance* that may be achieved, if the assumptions used in the Scharber model can be met. These assumptions implicitly incorporate a number of additional requirements – in particular related to the complicated bulk and interface behavior as well as to the exciton and charge-carrier dynamics – that have to be achieved in order to obtain a high-performance material. The standard parameters which reflect these assumptions (*e.g.*, the fill factor of 65%, the uniform external quantum efficiency of 65%, the required LUMO offset of 0.3 eV between donor and acceptor, and the empirical loss parameter of 0.3 eV) can in principle be further improved, but their practical realization already poses challenges. The PCE values reported in this paper correspond to a standard phenyl-C61-butyric acid methyl ester (PCBM) acceptor counterpart. While the Scharber model is clearly too simplistic to account for all the complex physics of an OPV, it nonetheless provides a valuable indication about the inherent promise of a candidate compound. A good PCE value is thus a necessary condition for a successful donor material (based on the principal energy levels of its molecular constituent), but not a sufficient one. It offers a guideline as to whether development efforts geared towards realizing the other material features have a chance of being worthwhile. However, there is no guarantee that the top candidates will indeed perform as well as indicated since they may fail for factors not captured in the employed analysis. Pharmaceutical screening efforts are a good analogy to the work presented here: our study reveals insights into new and potentially successful molecular motifs, which can then be further explored by experiment and more detailed calculations.

The theoretical PCE limit within the standard Scharber model with a PCBM acceptor is 11.1%, and it requires a LUMO energy of −4.00 eV and a gap of 1.41 eV (*i.e.*, the HOMO level has to be at −5.41 eV) as the optimum parameter combination.

Table 1 Statistical analysis of the calibration and averaging scheme employed in this study. The best estimates for the HOMO and LUMO values for each molecular motif (m-) incorporate up to 75 values from different model chemistries and geometries. When they are computed we obtain a mean absolute deviation (MAD) and root-mean-square deviation (RMSD) for each motif. We then analyze these statistical measures for the entire set of 2.3 million screened motifs (s-) and obtain the average (avg), MAD, and RMSD values given in the table rows.[28] All results are in eV

	HOMO		LUMO	
	m-MAD	m-RMSD	m-MAD	m-RMSD
s-avg	0.07	0.09	0.09	0.12
s-MAD	0.02	0.03	0.02	0.03
s-RMSD	0.06	0.25	0.07	0.26

Candidates with this HOMO level should also be stable towards oxidation in the air, as they remain below the threshold of −5.3 eV. Of the 2.3 million screened compounds, only about 1000 (*i.e.*, 0.04%) show a PCE of 11% and higher, and 35 000 (*i.e.*, 1.5%) a value of over 10%. Fig. 1(c) shows the predicted PCE distribution for the candidates. Most of them are estimated to have a PCE below 4%. The list of the top 10 000 candidates from our current analysis is provided in the ESI.†

The CEP results are also compiled in the Harvard Clean Energy Project Database (CEPDB),[30] which is made available as an open resource. It is designed for the storage and analysis of data on organic electronics. It allows to readily identify candidates with specific property combinations. The CEPDB also provides a simple interface to employ other, potentially more advanced device performance models. It will feature regularly updated lists of the most promising OPV candidates as more data becomes available, as we improve the calibration scheme, and augment the data analysis capability. The CEPDB was released to the public in June 2013 under a Creative Commons Attribution ShareAlike license.

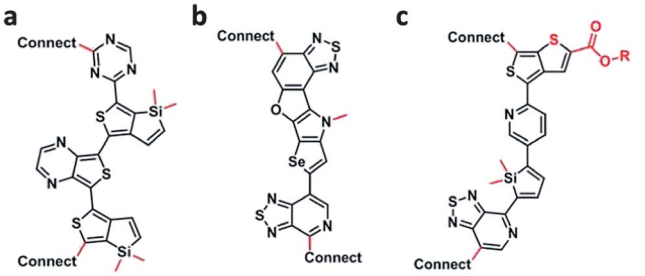

Fig. 2 Example structures from the top candidates list (each with potential modifications marked in red). (a) shows the candidate ranked #77 that has multiple inter-monomer nitrogen-sulfur interactions. These can facilitate a highly planarized structure in the solid-state, and thus potentially enhance the electronic coupling. (b) is the candidate ranked #5. It has a very rigid 5-ringed, heterocyclic co-monomer structure, which may reduce reorganization and relaxation energies. (c) displays the candidate ranked #1 that, after minor modifications marked in red, contains Yu's highly efficient thienothiophene co-monomer.[31] This co-monomer has been utilized in organic photovoltaic materials that have consistently surpassed 7.0% power conversion efficiency.

C. Empirical assessment of the top candidates

A visual survey of the highest ranked candidates (examples are shown in Fig. 2) immediately reveals the prevalence of moieties (*e.g.*, benzothiadiazole), which are known to be extremely useful in polymeric OPV materials and which appear in numerous high-performance devices reported in the literature. The candidates also possess attributes thought to be imperative for efficient electronic materials: The inherent stiffness of fused multi-ring systems lends itself to reducing energy loss mechanisms by minimizing reorganization and relaxation energies. The heteroatom substitutions on the periphery of these ring systems may allow for through-space interactions that could enhance electronic coupling and provide morphological control during device fabrication. Loss mechanisms and the lack of morphological control can result in poor performance of even exceptional materials. These empirical observations lend confidence to our predictions. The top candidates feature some relatively complicated multi-ring patterns and tetramer repeat units, which are seldomly approached by experimentalists. Although fused-ring systems have been reported, multi-heteroatom substituted ring structures and tetramer repeat units are very rare due to synthetic challenges (most OPV polymers have trimer repeat units). Advances in the available synthetic tools justify the hope that the reported lead candidates may broaden the range of experimental target compounds.

D. Structural analysis of the top candidates

Following the ranking of the molecular motifs and their empirical assessment, we now analyze the structural composition of the top candidates (*i.e.*, the list of candidates with a PCE of more than 10%) more systematically. Our goal is to identify patterns and hints towards structure–property relationships in these most promising compounds.[32]

In a first step we perform a statistical analysis with respect to the occurrence of the molecular building blocks that were used

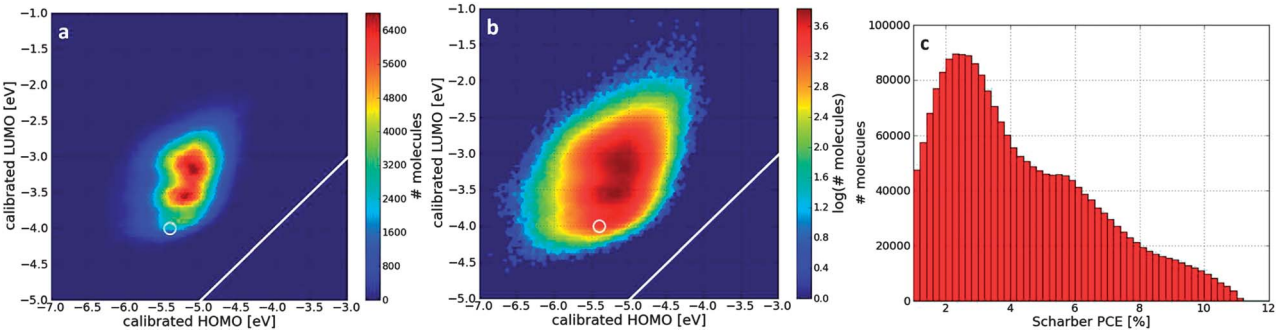

Fig. 1 Screening results and analysis. Dynamic gap range plot of the 2.3 million molecular motifs screened to date on a linear (a) and logarithmic (b) scale. The available combinations of HOMO and LUMO values span a wide range, as is particularly apparent in (b). However, the parameter space for high-performance materials (marked by the white circle) is small and only contains a relatively sparse distribution of candidates. It is worth noting that the screened structures cluster around LUMO values that would be more suitable for different acceptor materials, but that the gap of many candidates is too large for the efficient harvesting of the solar spectrum. Panel (c) shows the resulting power conversion efficiency (PCE) histogram according to the Scharber model with respect to a phenyl-C61-butyric acid methyl ester (PCBM) acceptor.[22] Candidates without photovoltaic activity (*i.e.*, PCE = 0%[29]) have been excluded in this graph.

in the library generation. We employ a hypergeometric distribution analysis to assess the prevalence of the 26 basic fragments in the top candidates.[33] The hypergeometric distribution gives the probability of finding k observations in a subpopulation of size n, given that there are K observations in the entire population of size N. This translates to counting, for each building block, the number of molecular motifs that contain it in the high-PCE subset and in the entire library, respectively. Enrichment or depletion relative to a random distribution is reflected in the resulting Z-score of each building block. The Z-score is given by $z = (k - \langle k \rangle)/\sigma(k)$. Here, k is the observed number of molecules in the top set containing the building block of interest; $\langle k \rangle$ is its expectation value given by $\langle k \rangle = n \cdot K/N$; and $\sigma(k)$ is the standard deviation of the hypergeometric distribution. A positive Z-score indicates that a building block is found more often than is statistically expected in a random distribution, and a negative Z-score registers the opposite. We find that certain building blocks are very common in the top candidates while others are underrepresented. Fig. 3 and Table 2 show these findings in detail. We conclude that the performance of the best candidates can at least in part be traced back to their structural composition.

The following building blocks (termed enriched monomer motifs) are strongly overexpressed in the top PCE set: [1,2,5]-thiadiazolo[3,4-*C*]pyridine (**26**), pyridine (**17**), benzothiadiazole (**15**), silacyclopenta-2,3-diene (**3**), and 2*H*-2-silaindene (**24**). Other building blocks (depleted monomer motifs) are significantly underrepresented. The five most depleted fragments are pyrrole (**5**), isoindole (**22**), cyclopentadiene (**7**), 1*H*-thieno[3,4-*b*]-pyrrole (**11**), and isoindene (**25**). All the basic building blocks are shown in Fig. 4.

There has been much work related to successful OPV donor compounds based on thiadiazole moieties, such as **15** and **26**, and they have been popular in the development of OPV donor materials for some time. Our findings about the inherent promise of these two building blocks are thus supported by a substantial amount of empirical evidence. Thiadiazole moieties can act as electron-withdrawing groups in co-polymers. Benzothiadiazole (**15**) for instance has successfully been coupled with electron-donating moieties (*e.g.*, fluorene and carbazoles) to lower the donor bandgap.[34–37] Following this approach, Blouin *et al.*[38] devised a small molecular library from which the thiadiazolo-pyridine co-polymer emerged with a very promising Scharber PCE. This finding gives an interesting context to the observed prevalence of **17** (in addition to **15** and **26**) in our top candidates. Recently, there have also been developments towards using the combination of thiadiazole and pyridine in optoelectronic materials.[39,40] The thiazole unit (**19**) is an

Table 2 *Z*-scores of the most amplified and depleted fragments for the lists of candidates with power conversion efficiency (PCE) ≥ 10%. The other columns show the mean and median PCE value (in %) for all motifs containing the respective building block

Building Block ID	*Z*-score PCE ≥ 10%	Mean PCE	Median PCE
26	438.5	6.3	8.1
17	138.5	4.4	3.8
15	62.2	5.8	5.8
3	34.6	4.7	4.5
24	34.6	5.5	5.6
5	−123.6	3.0	2.4
22	−80.3	2.7	2.2
7	−60.2	3.7	3.3
11	−53.1	2.7	2.3
25	−51.4	3.8	3.5

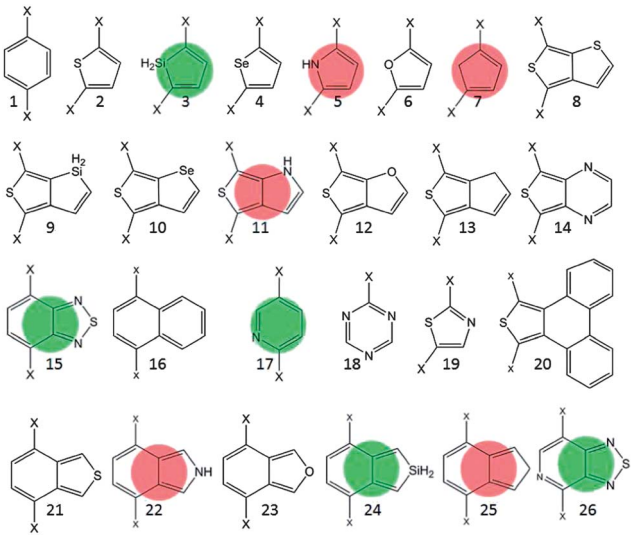

Fig. 4 The 26 building blocks (including chemical handles indicated by 'X') of the molecular candidate library. Moieties with the most amplified occurrence in the top candidates (relative to the statistical expectation) are highlighted in green, and red indicates the ones with the most decreased occurrence.

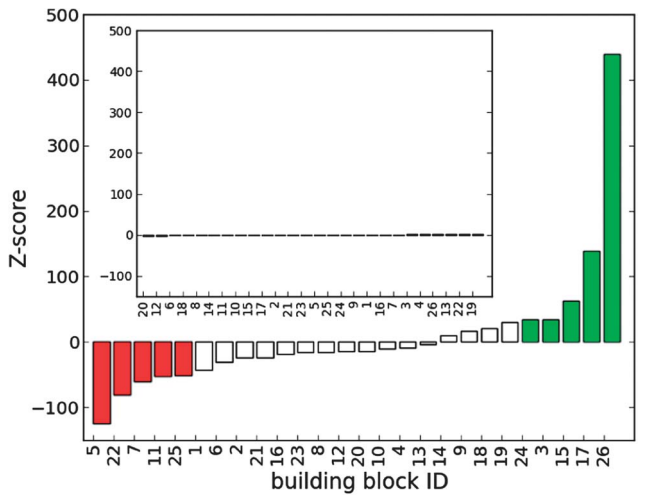

Fig. 3 *Z*-scores of the top candidates with power conversion efficiency (PCE) ≥ 10%. We see a number of building blocks with elevated appearance while others are depleted. The inlay shows a random sample, which does not exhibit any significant enrichment or depletion. The PCE ≥ 11% list gives qualitatively the same result as the one for PCE ≥ 10%, *i.e.*, this analysis of the high-performance candidates appears to be robust to the cutoff.

electron-withdrawing group as well, and it has been employed in the scaffold of semiconductor materials for both photovoltaics[41,42] and light-emitting diodes.[43] Si-containing building blocks – 2*H*-2-silaindene (**24**) and silacyclopenta-2,3-diene (**3**) – are also amongst the five monomer motifs with the largest *Z*-scores. It was shown that modifying an electron-donating moiety such as fluorene to a silafluorene can successfully increase the photovoltaic efficiency of an organic semiconductor.[44] Recent work along those lines has focused on benzosiloles.[45,46] It has also been suggested that silafluorenes provide an advantage in solution processing. Corey notes that despite these successes and the apparent potential of Si-based moities, the application of silaindenes (such as **24**) has not been pursued in the development of new OPV materials.[47] The fact that **24** is notably overrepresented in our top candidates underscores that these moieties hold great prospects.

We also mention the five fragments that have the most negative *Z*-score, which means that finding them in a top compound with a large PCE is much less likely than would be statistically expected. Two of these (isoindole and isoindene, *i.e.*, **22** and **25**) are analogous to the high *Z*-score silaindene moiety mentioned above. Their lack of promise compared to the Si-analogue further supports our conjecture concerning the untapped potential of Si-heterocyclic OPV designs. The simple fragments pyrrole (**5**) and cyclopentadiene (**7**) also show less promise than most of the other building blocks. The latter was only chosen for a homologue comparison in the first place.

E. Correlations between building blocks and performance

After having identified the structural patterns that can be linked to the most promising candidates, we now address the question of how decisive these individual building blocks are for the overall performance of a candidate. For that we compute the mean and median PCE of all compounds that contain a particular fragment.[29] Since a compound is constructed from several fragments with different *Z*-scores, the effect of each individual fragment will in principle average out. However, if the influence of a building block is dominant for the overall performance, then we should still find a discernible structure in the results. In a follow-up study we will address the question of moiety combinations and the impact of their joint occurrence, as well as that of generalized structural patterns.

As can be seen in Fig. 5 and Table 2, the assessment of the fragment quality based on the *Z*-scores is essentially confirmed by the PCE statistics. The enriched monomer motifs are associated with higher PCEs compared to most other fragments, while the depleted monomer motifs underperform significantly. **17** shows a lower PCE than anticipated based on its *Z*-score, which suggests that its prevalence in the high-performance candidates is conditional and subject to being paired with certain conterparts. This interpretation aligns very well with the design of push-pull-copolymers as discussed above.[38] Two additional fragments of interest emerge from the PCE analysis, *i.e.*, thieno[3,4-*b*]pyrazine (**14**) and 1,3,5-triazine (**18**). Candidates based on either of these two fragments tend to perform very solidly, but their modest *Z*-scores indicate that this

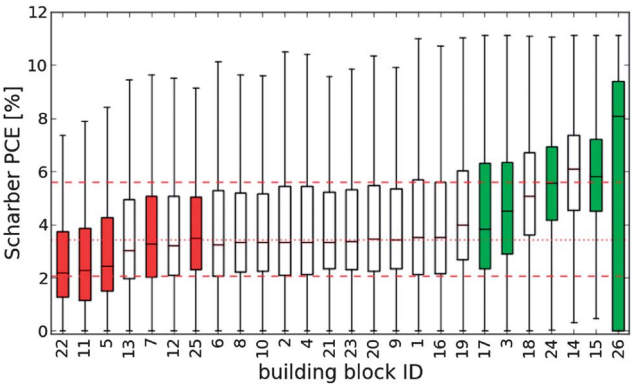

Fig. 5 Box-and-whisker plot for the distribution of Scharber power conversion efficiency (PCE) values associated with each library building block. The moieties are sorted by their mean PCE. The dotted and dashed red lines represent the median PCE and the 25/75th percentile over all candidates, respectively. The limits of each box represent the 25/75th percentile distribution for molecules that contain the corresponding building block, and the median value is shown as a black line. The whiskers extend to the extreme data points. The moieties with the five highest and lowest *Z*-scores are represented in green and red, respectively.

does not extend into the highest PCE region. All fragments with the exception of **26** show a relatively narrow PCE window into which the majority of the derived candidates fall. **26** in contrast covers a much larger variance. That means, that its performance is essentially hit-or-miss. It is remarkable that the 75th percentile is over 9% PCE, and the 50th percentile is over 8%, but then the 25th percentile is at 0% PCE. This feature can probably be traced back to the fact that a LUMO offset of $\geq$ 0.3 eV is required in the Scharber model. Compounds that do not fulfill this requirement are predicted to show no charge separation and their PCE drops to 0%. Finally, it also emerges that a majority of fragments never reach the theoretical maximum of the Scharber model.

III. Conclusions

We can conclude that the Harvard Clean Energy Project with its virtual, high-throughput, *first-principles* quantum chemical characterization of candidate compounds and its *big data* approach provides a framework for the rational, systematic, and accelerated development of new organic electronic materials. We present the top candidates that have emerged after investigating 2.3 million molecular motifs and performing 150 million DFT calculations. By analysing these and the candidate pool in its entirety, we identified building blocks such as thiadiazoles and silaindenes, that are related to the ideal energy level alignment for high-performance OPV donor materials. Consequently, these are of particular interest for the design of future materials, while less promising fragments may not have to be considered in future searches. These insights will aid in the transition from a brute-force screening approach towards the active design and engineering of new molecular materials.

In upcoming publications, we will discuss other correlations between the top candidates and their topological and physicochemical features. We will utilize our insights into the

underlying structure–property relationships in the creation of second generation screening libraries and as a starting point for the construction of new candidates *via* genetic algorithms.[48–50] In addition to expanding and improving our candidate characterization and data analysis capability, we will also employ other OPV performance models in order to advance the quality and robustness of our predictions. Finally, we will generalize our work to materials for multi-junction devices. In the spirit of open science we have made the CEPDB available to the public and hope that other research groups will use the released data for their own scientific pursuits.

Author contributions

AAG conceived the Clean Energy Project. AAG and JH designed the research, and JH was responsible for the implementation and deployment of the virtual screening platform, database, result management and processing, Scharber analysis, and ranking. ROA and AJ carried out the structural analysis of the top candidates with contributions from MBF. ALA and ZB performed their empirical assessment. ROA, ALA, RM, AS, and ZB designed and ROA implemented the primary screening library. ROA was also responsible for the submission of the workunits to the grid, supported by SS and JH. LRS, SAE, JH, and SE contributed to the calibration scheme. CRS, KT, and JH were responsible for the public release of the database. JH and ROA wrote the manuscript with contributions from all other authors.

Acknowledgements

We wish to thank IBM for organizing the World Community Grid and the WCG members for their computing time donations. Additional information about the project as well as links to join the WCG and the CEP can be found in ref. 7 and 51. We are grateful for discussions with and input from Carlos Amador-Bedolla (UNAM Mexico); Xueliang Liu, Aidan Daly, and Jarrod McClean (Harvard University); Michael Toney, Arjan Zoombelt, Jianguo Mei, Yan Zhou, and Alex Ayzner (Stanford University); Tim Mueller and André Botelho (Johns Hopkins University); Markus Scharber (University of Linz); Mark Schofield (Haverford College); Sergei Tretiak (LANL); Troy Van Voorhis (MIT); Marcus Hanwell (Kitware Inc.); Björn Zimmermann (Wolfram Research); as well as for support by the IBM WCG team, the WCG Community Advisors, the Q-Chem development team, and Harvard FAS Research Computing. We acknowledge support from Cyrus Wadia and the Materials Genome Initiative.[52] The CEP has been mainly funded by the U.S. Department of Energy through grant no. DE-SC0008733 and 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). We acknowledge additional support from the U.S. National Science Foundation under grants no. DMR-0820484 and DMR-0934480. ROA was supported by a Giorgio Ruffolo Fellowship in the Sustainability Science Program at Harvard University's Center for International Development. MBF acknowledges support by the U.S. Department of Energy, Office of Science Graduate Fellowship Program, administered by ORISE-ORAU under contract no. DE-AC05-06OR23100. SE performed work as part of the Fellowships for Young Energy Scientists program of the Foundation for Fundamental Research on Matter (FOM), which is part of the Netherlands Organization for Scientific Research (NWO). We acknowledge software support from Q-Chem Inc., ChemAxon Ltd., Kitware Inc., Molecular Networks GmbH, and Optibrium Ltd. Additional computing time was provided on the Odyssey cluster supported by the Harvard FAS Research Computing Group; the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by U.S. National Science Foundation grant no. OCI-1053575; the Center for Functional Nanomaterials, Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under contract no. DE-AC02-98CH10886; and Cycle Computing.

References

1 A. Heeger, in *Global Sustainability - A Nobel Cause*, ed. H. J. Schellnhuber, M. Molina, N. Stern, V. Huber and S. Kadner, Cambridge University Press, Cambridge, UK, 2010.
2 S. R. Forrest, *Nature*, 2004, **428**, 911–918.
3 M. S. Dresselhaus, G. Dresselhaus and P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes: Their Properties and Applications*, Academic Press, San Diego, 1996.
4 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
5 R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt and A. Aspuru-Guzik, *Energy Environ. Sci.*, 2011, **4**, 4849–4861.
6 C. Amador-Bedolla, R. Olivares-Amaya, J. Hachmann and A. Aspuru-Guzik, in *Informatics for Materials Science and Engineering*, ed. K. Rajan, Elsevier, Amsterdam, 2013.
7 http://www.worldcommunitygrid.org, accessed 3 July 2013.
8 D. Clery, *Science*, 2005, **308**, 773.
9 N. M. O'Boyle, C. M. Campbell and G. R. Hutchison, *J. Phys. Chem. C*, 2011, **115**, 16200–16210.
10 I. Y. Kanal, S. G. Owens, J. S. Bechtel and G. R. Hutchison, *J. Phys. Chem. Lett.*, 2013, **4**, 1613–1623.
11 N. Bérubé, V. Gosselin, J. Gaudreau and M. Côté, *J. Phys. Chem. C*, 2013, **117**, 7964–7972.
12 Y. Zhang, W. Shen, R. He, X. Liu and M. Li, *J. Mater. Sci.*, 2012, **48**, 1205–1213.
13 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191–201.
14 G. L. W. Hart, *Nature*, 2012, **491**, 674–675.
15 W. F. Maier, K. Stöwe and S. Sieg, *Angew. Chem., Int. Ed.*, 2007, **46**, 6016–6067.

16 A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2011, **50**, 2295–2310.
17 S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko and D. Morgan, *Comput. Mater. Sci.*, 2012, **58**, 218–226.
18 D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov and K. W. Jacobsen, *Comput. Sci. Eng.*, 2012, **14**, 51–57.
19 L. Yu and A. Zunger, *Phys. Rev. Lett.*, 2012, **108**, 068701.
20 J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, *Nat. Chem.*, 2009, **1**, 37–46.
21 In our previous paper we used the term 'connectivities' for 'molecular motifs' and 'molecular motifs' for 'geometries', but believe that the terminology chosen here is more suitable.
22 M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, A. J. Heeger, C. Waldauf and C. J. Brabec, *Adv. Mater.*, 2006, **18**, 789–794.
23 T. Ameri, G. Dennler, C. Lungenschmied and C. J. Brabec, *Energy Environ. Sci.*, 2009, **2**, 347–363.
24 W. Shockley and H. J. Queisser, *J. Appl. Phys.*, 1961, **32**, 510–519.
25 S. Hamel, P. Duffy, M. E. Casida and D. R. Salahub, *J. Electron Spectrosc. Relat. Phenom.*, 2002, **123**, 345–363.
26 J. Luo, Z. Q. Xue, W. M. Liu, J. L. Wu and Z. Q. Yang, *J. Phys. Chem. A*, 2006, **110**, 12005–12009.
27 G. Zhang and C. B. Musgrave, *J. Phys. Chem. A*, 2007, **111**, 1554–1561.
28 In the current analysis we have not removed duplicate geometries (which can occur if the DFT optimization collapses different conformer guesses to the same minimum), the results of unrestricted calculations that do not show spin polarization, or pathologic cases.
29 The PCE in the Scharber model can become negative if the HOMO level of the donor is very similar to the LUMO of the acceptor. In that case, however, the Scharber model is not well defined and a candidate is no OPV. We thus set all negative PCE values to zero, and PCE = 0% is used in the calculation of the mean PCE values.
30 http://cepdb.molecularspace.org, accessed 3 July, 2013.
31 H.-Y. Chen, J. Hou, S. Zhang, Y. Liang, G. Yang, Y. Yang, L. Yu, Y. Wu and G. Li, *Nat. Photonics*, 2009, **3**, 649–653.
32 M. Misra, D. Andrienko, B. Baumeier, J.-L. Faulon and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2011, **7**, 2549–2555.
33 J. A. Rice, *Mathematical Statistics and Data Analysis*, Duxbury Press, Pacific Grove, CA, 3rd edn, 2006.
34 M. Svensson, F. Zhang, S. Veenstra, W. Verhees, J. Hummelen, J. Kroon, O. Inganäs and M. Andersson, *Adv. Mater.*, 2003, **15**, 988–991.
35 O. Inganäs, M. Svensson, A. Gadisa, F. Zhang, N. Persson, X. Wang and M. Andersson, *Appl. Phys. A: Mater. Sci. Process.*, 2004, **79**, 31–35.
36 D. Mühlbacher, M. Scharber, M. Morana, Z. Zhu, D. Waller, R. Gaudiana and C. Brabec, *Adv. Mater.*, 2006, **18**, 2884–2889.
37 J. Chen and Y. Cao, *Acc. Chem. Res.*, 2009, **42**, 1709–1718.
38 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.
39 Y. Sun, S.-C. Chien, H.-L. Yip, Y. Zhang, K.-S. Chen, D. F. Zeigler, F.-C. Chen, B. Lin and A. K.-Y. Jen, *J. Mater. Chem.*, 2011, **21**, 13247–13255.
40 X. Yong and J. Zhang, *J. Mater. Chem.*, 2011, **21**, 11159–11166.
41 T. Ro and J. Hong, *Bull. Korean Chem. Soc.*, 2012, **33**, 2897–2902.
42 S. V. Mierloo, A. Hadipour, M.-J. Spijkman, N. V. D. Brande, B. Ruttens, J. Kesters, J. D. Haen, G. V. Assche, D. M. D. Leeuw, T. Aernouts, J. Manca, L. Lutsen, D. J. Vanderzande and W. Maes, *Chem. Mater.*, 2012, **24**, 587–593.
43 S. Ando, J. Nishida, H. Tada, Y. Inoue, S. Tokito and Y. Yamashita, *J. Am. Chem. Soc.*, 2005, **127**, 5336–5337.
44 E. Wang, L. Wang, L. Lan, C. Luo, W. Zhuang, J. Peng and Y. Cao, *Appl. Phys. Lett.*, 2008, **92**, 033307.
45 R. V. Solomon, A. P. Bella, S. A. Vedha and P. Venuvanalingam, *Phys. Chem. Chem. Phys.*, 2012, **14**, 14229–14237.
46 M. Yuan, P. Yang, M. M. Durban and C. K. Luscombe, *Macromolecules*, 2012, **45**, 5934–5940.
47 J. Y. Corey, *Adv. Organomet. Chem.*, 2011, **59**, 181–328.
48 W. Paszkowicz, *Mater. Manuf. Processes*, 2009, **24**, 174–197.
49 V. Arora and A. Bakhshi, *Chem. Phys.*, 2010, **373**, 307–312.
50 R. Giro, M. Cyrillo and D. S. Galvão, *Chem. Phys. Lett.*, 2002, **366**, 170–175.
51 http://cleanenergy.molecularspace.org, accessed 3 July 2013.
52 http://www.whitehouse.gov/mgi, accessed 3 July 2013.