SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Energy

» Solar cells

» Electronic structure

Received: 02 December 2015
Accepted: 12 August 2016
Published: 27 September 2016

Data Descriptor: The Harvard organic photovoltaic dataset

Steven A. Lopez[1,*], Edward O. Pyzer-Knapp[1,*], Gregor N. Simm[1], Trevor Lutzow[1], Kewei Li[1], Laszlo R. Seress[1], Johannes Hachmann[2,3,4] & Alán Aspuru-Guzik[1]

The Harvard Organic Photovoltaic Dataset (HOPV15) presented in this work is a collation of experimental photovoltaic data from the literature, and corresponding quantum-chemical calculations performed over a range of conformers, each with quantum chemical results using a variety of density functionals and basis sets. It is anticipated that this dataset will be of use in both relating electronic structure calculations to experimental observations through the generation of calibration schemes, as well as for the creation of new semi-empirical methods and the benchmarking of current and future model chemistries for organic electronic applications.

Design Type	data integration objective • database creation objective
Measurement Type(s)	molecular orbitals and OPV bulk properties
Technology Type(s)	data item extraction from journal article
Factor Type(s)	
Sample Characteristic(s)	

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260, USA. [3]Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, New York 14260, USA. [4]New York State Center of Excellence in Materials Informatics, Buffalo, New York 14203, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.A.-G. (email: aspuru@chemistry.harvard.edu).

Background & Summary

Standard data sets used for the calibration of computational results have been extremely useful for the development of electronic structure methods and their application to areas such as thermochemistry[1–3] as well as non-covalent interactions[4,5]. To our knowledge, the field of organic photovoltaics, as it pertains to high-throughput virtual screening[6–8], lacks a similar collection of data. Since the relationship between theoretically predicted and experimentally observed properties is often non-trivial, the dissemination of directly comparable data for a well-defined set of molecules can be a great asset to accelerate advances in this field.

Many areas of materials chemistry have benefited from the application of high-throughput virtual screening, which has led to an accelerated discovery of new materials[6–16]. Since this approach allows a large number of compounds and materials to be pre-screened using efficient *in silico* (often quantum-chemical) techniques, it allows experimental scientists to focus time and resources on fewer, more promising, candidates[17]. However, theoretical studies (*i.e.*, based on density functional theory (DFT)) only approximates the observed experimental properties and care must be taken when relating one to the other[18]. The Scharber model[19] is utilized to compute the maximum percent conversion efficiencies for the 350 studied molecules. The quantities that enter the Scharber model are the lowest unoccupied molecular orbital (LUMO) and highest occupied molecular orbital (HOMO) energies energy and the HOMO-LUMO gap. These are used to compute the open circuit potential (V_{OC}) and short circuit current density (J_{SC}). Percent conversion efficiency (PCE) is the computed according to equation 1.

$$\mathrm{PCE} = 100 * \frac{V_{OC} * \mathrm{FF} * J_{SS}}{\mathrm{Pin}} \quad (1)$$

In the Scharber model, the fill factor (FF) is set to 65%, and *J*sc is qualitatively related to the HOMO-LUMO gap.

One area in which this method has been most visibly applied is the area of organic photovoltaic materials[6,7], with the Harvard Clean Energy project being an example[8,20]. Many approximations are made to efficiently screen of millions of compounds. An ability to relate these calculations to experimental data is critical for the implementation of an efficient feedback loop. We believe that such a feedback loop is vital for the ongoing success of collaborative efforts. Unfortunately, there are very few collections of experimental results from which to build these models, and we are not aware of any significantly sized set of molecules for which both quantum chemical and experimental values are reported.

Here we report the Harvard Organic Photovoltaic Dataset (HOPV15) consisting of both experimental results compiled from the literature, and corresponding data from quantum chemical calculations using a selection of five functionals chosen to contain both generalized-gradient approximation (BP86 (refs 21,22) and hybrid designs with a range of incorporated amounts of exact exchange PBE0 (refs 23,24), B3LYP[21,25], and M06-2X[26,27] in combination with the double-ζ def2-SVP basis set[28]. It will have a multitude of uses, including the calibration of quantum chemical results to experimental observables[29], the development of new methodologies for property estimation[19], as well as the design of new Hamiltonians for semi-empirical methods[30].

The compounds in this data set represent a diverse cross-section of molecular designs in this field. This is reflected in the Tanimoto distance between each molecule and all others as described by the 512-bit, radius-2 Morgan circular fingerprint[31]. We only calculate the upper triangular of the distance matrix. A histogram of the computed distances is shown in Fig. 1 and it emphasizes that the average Tanimoto distance is just below 0.8 (the Tanimoto distance is bounded at 0 for a perfect match between fingerprints and 1.0 for no common bits in the fingerprint). An average distance of 0.8, therefore, is a good indication of the molecular diversity of this data set.

Methods

Simplification of the Conformational Energy Landscape

Many of the structures reported in the literature had long alkyl chains added to the 'active' photovoltaic core in order to improve solution processing. This substantially convolutes the conformational energy landscape, while the electronic structure are not significantly changed. This was confirmed by studying the effect of chain length on the HOMO-LUMO gap of these molecules with a) the original chain length, b) the chain reduced to two carbons c) the chain length reduced to one carbon and d) the chain removed entirely. It was observed that there was no significant difference on the HOMO-LUMO gap when the chain was reduced to one or two carbons ($\Delta\mathrm{Gap} = 0.0 \pm 0.01$ eV), and a small difference when the chain was removed altogether ($\Delta\mathrm{Gap} = 0.05 \pm 0.08$ eV). Since a complicated conformational landscape necessitates the generation of an exponentially growing number of conformers (the number of conformers scales approximately as 3^N where N is the number of rotatable bonds) and can thus reduce the performance of many of the common conformer generation algorithms, we decided to truncate alkyl chains to a methyl group.

Generation of Molecular Conformations

Starting from the simplified molecular-input line-entry system (SMILES[32]) string representation of the molecule, with all alkyl chains reduced to one carbon, 1500 initial guesses at the 3D conformation of the molecule were generated using the conformer generation package included in the open-source RDKit

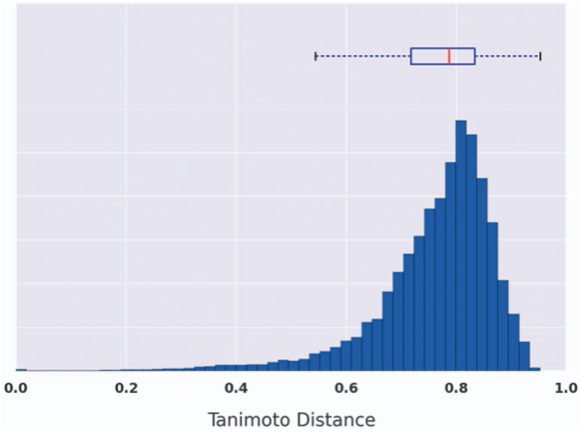

Figure 1. The distribution of Tanimoto distances in the distance matrix calculated for the Harvard Organic Photovoltaic 2015 dataset presented in this work suggests that the data set encapsulates significant molecular diversity. The box-plot above the histogram shows the mean, 25 and 75% percentile values with 10 and 90% points indicated by the whiskers.

software[33]. These initial guesses were then minimized using the MMFF force field[34] implemented in this package[35], with duplicate structures resulting from initial guesses minimizing to the same relaxed structure removed using the *obfit* functionality implemented in the *OpenBabel* software package[36]. The lowest energy conformation from each of up to twenty clusters which fell within a window of 5 kcal mol^{-1} were selected to represent energetically feasible conformations for the molecule which may contribute to the performance of the material, especially in disordered or semi-ordered materials.

Quantum-chemical Calculations

The geometries for every selected conformation were minimized using the BP86 functional, and the def3-SVP basis set. For force-field minimizations, duplicate structures resulting from multiple force field minima converging to the same BP86/def2-SVP minimum were removed using the *obfit* functionality implemented in OpenBabel software package[36] with a tolerated RMSD in atomic positions of >0.1 Å.

For each unique conformation, single point energies were calculated with PBE0 (refs 23,24), B3LYP[21,25], and M06-2X[26,27] in combination with the double-ζ def2-SVP basis set[28]. As previously stated, these functionals represent a range of exact exchange, with BP86 (refs 21,22) (0%) and M06-2X[26,27] (52%) representing the extremes of the range. The inclusion of a range of functionals increases the utility of the data set since the additional information can be used to either benchmark performance against a range of model chemistries, or alternatively these results can be used in an *ensemble average* to provide a model, which is more general than any individual model chemistry.

Code availability

Quantum-chemical calculations were performed using Q-Chem version 4.1.2, and is available from http://www.q-chem.com under a commercial license.

The OpenBabel software package is freely available from http://openbabel.org/ under the GPL license[37].

The RDKit is freely available from http://rdkit.org/ under the BSD licence[38].

Data Records

The data set is shared publically on *Figshare* (Data Citation 1). We extensively searched the literature and located 350 small molecules and polymers that were utilized as *p*-type materials in OPVs. For each reported molecule, atomic coordinates, experimental properties and their calculated equivalents are stored in a plain-text XYZ-format described below. Deposited are the 350 molecules which make up the HOPV dataset, up to twenty of their low-energy calculated molecule conformations and, where available, the power conversion efficiency (PCE), open circuit potential (V_{OC}), short circuit current density (J_{SC}), highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy and the HOMO-LUMO gap. The reported PCE, V_{OC}, J_{SC}, HOMOs, and LUMOs are reported in percent, Volts, mA/cm^2, and atomic units, respectively.

File format

Figure 2 shows the makeup of the HOPV_15.data file (per molecule). It is an extension of the commonly used XYZ format for encoding Cartesian coordinates of molecules, with no formal specification. It

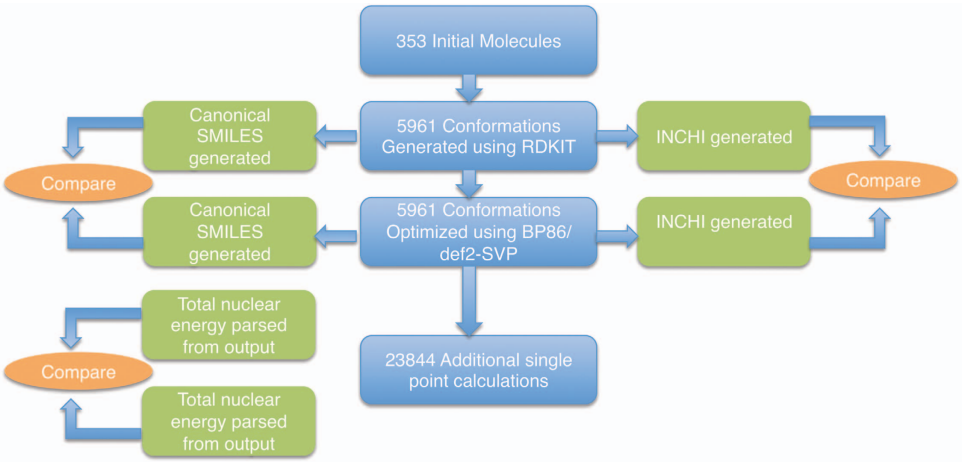

Figure 2. The workflow for validating calculated geometries by comparing the canonical SMILES representation and the InChI representation generated for the initial force field optimized geometry for a conformer and the DFT optimized geometry for that conformer. The total nuclear energy for each conformer was calculated and compared to that reported in the related quantum-chemical output files.

Line	Contents
1	SMILES of molecule
2	InChI of molecule
3	Experimental data (as CSV)
4	'Pruned' smiles of molecule
5	Total number of conformers
	For each conformer (N = index, n = number of atoms)
L = 5+(N *n)	Conformer number
L+1	Number of atoms
L2 = L+2-> L+2+(N*n)	Atomic element, X, Y, Z coordinates
	For each functional:
L2+1-> L2+5	Calculated Data (as CSV)

Table 1. A description of the file-format used in the HOPV15 data file.

contains a header line specifying the number of atoms n, a comment line, and n lines containing element type and atomic coordinates, one atom per line. We have extended this format as indicated in Table 1 in a manner similar to Von Lilienfeld *et al.*[39] did for purposes of machine learning. In addition to the XYZ format for the storing of electronic coordinates, experimental and calculated properties are stored in CSV format as described in Tables 2 and 3, respectively.

Technical Validation

Validation of Computational Results

For each conformer generated from the initial molecules, a DFT optimized geometry was generated at the BP86 (refs 21,22)/def2-SVP[28] level of theory. Geometries were validated using a technique similar to that used in the work of von Lilienfeld *et al.*[40] To detect instances where the DFT optimized geometry had changed drastically from the initial force field optimized geometry, both the InChI and canonical SMILES were generated for both geometries, and compared. The InChI and canonical SMILES are both theoretically unique identifiers, and so comparing these two descriptors represents a method for evaluating if the minimized structure consistent with typical geometries.

In order to validate the computation of the electronic structure of the conformations, an additional test was performed on all optimizations, and single point electronic structure calculations. The total nuclear energy, a property, which is solely reliant on the nuclear positions and charges, was calculated for each conformation. This was then compared to the reported values within the Q-Chem[41] output files for each calculation. This technique has been utilized as part of the Harvard Clean Energy Project's validation suite for calculations performed on the World Community Grid[42] and is aimed at testing for hardware

1	Digital Object Identifier
2	InChIKEY of molecule
3	Construction (Polymer/molecule)
4	Architecture
5	Complement
6	HOMO
7	LUMO
8	Electochemical gap
9	Optical gap
10	PCE
11	V_{OC}
12	J_{SC}
13	Fill factor

Table 2. A description of the CSV format for storing experimental information.

Index	Contents
1	Functional/Basis set description
2	HOMO
3	LUMO
4	Gap
5	Scharber PCE
6	Scharber V_{OC}
7	Scharber J_{SC}

Table 3. A description of the CSV format used to store calculated properties.

issues which may negatively influence the quality of the calculated result. All calculations on molecules within the HOPV15 dataset passed both of these tests, which demonstrates the validity of the data set.

Validation of Experimental Results

The experimental results contained within this data set are taken from the literature, and so have been validated using the peer-review system. Wherever possible, molecules were taken from reviews, and cross referenced against the original publication to reduce the potential for transcription errors. In this way, the choice of molecules and the quality of the data has been validated by an external scientist (the composer of the review) who is also a domain expert. Where multiple reports for the same architecture exist, the most recent value was taken.

References

1. Curtiss, L. A., Raghavachari, K., Redfern, P. C. & Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **106,** 1063–1079 (1997).
2. Curtiss, L. A., Raghavachari, K., Trucks, G. W. & Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *J. Chem. Phys.* **94,** 7221–7230 (1991).
3. Amir Karton, S. D. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **510,** 165–178 (2011).
4. Jurečka, P., Šponer, J., Černý, J. & Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **8,** 1985–1993 (2006).
5. Řezáč, J., Riley, K. E. & Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **7,** 2427–2438 (2011).
6. Kanal, I. Y., Owens, S. G., Bechtel, J. S. & Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **4,** 1613–1623 (2013).
7. O'Boyle, N. M., Campbell, C. M. & Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **115,** 16200–16210 (2011).
8. Hachmann, J. *et al.* Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Env. Sci* **7,** 698 (2014).
9. Huskinson, B. *et al.* A metal-free organic–inorganic aqueous flow battery. *Nature* **505,** 195–198 (2014).
10. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater* **1,** 011002 (2013).
11. Shu, Y. & Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **142,** 104104 (2015).
12. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12,** 191–201 (2013).

13. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem* **4,** 83–89 (2012).
14. Colón, Y. J., Fairen-Jimenez, D., Wilmer, C. E. & Snurr, R. Q. High-Throughput Screening of Porous Crystalline Materials for Hydrogen Storage Capacity near Room Temperature. *J. Phys. Chem. C* **118,** 5383–5389 (2014).
15. Halls, M. D. & Tasaki, K. High-throughput quantum chemistry and virtual screening for lithium ion battery electrolyte additives. *J. Power Sources* **195,** 1472–1478 (2010).
16. Halls, M. D., Giesen, D. J., Hughes, T. F., Goldberg, A. & Cao, Y. *High-throughput quantum chemistry and virtual screening for OLED material components*, in **8829,** 882926–882926 (2013).
17. Pyzer-Knapp, E. O., Suh, C., Gomez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **45,** 195–216 (2015).
18. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for Density Functional Theory. *Chem Rev* **112,** 289–320 (2012).
19. Scharber, M. C. *et al.* Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Adv. Mater.* **18,** 789–794 (2006).
20. Hachmann, J. *et al.* The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J Phys Chem Lett* **2,** 2241–2251 (2011).
21. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A* **38,** 3098–3100 (1988).
22. Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys Rev B* **33,** 8822–8824 (1986).
23. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77,** 3865–3868 (1996).
24. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105,** 9982–9985 (1996).
25. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).
26. Zhao, Y. & Truhlar, D. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120,** 215–241 (2008).
27. Zhao, Y. & Truhlar, D. G. Density functionals for noncovalent interaction energies of biological importance. *J. Chem. Theory Comput.* **3,** 289–300 (2007).
28. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **7,** 3297–3305 (2005).
29. Pyzer-Knapp, E. O., Simm, G.N. & Aspuru-Guzik, A. Bayesian Calibration of Quantum Chemical Calculations to Experimental Observations: Application to Organic Photovoltaics. *arXiv* **1510,** 00388.
30. Botelho, A. L., Shin, Y., Liu, J. & Lin, X. Structure and Optical Bandgap Relationship of π-Conjugated Systems. *PLoS ONE* **9,** e86370 (2014).
31. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50,** 742–754 (2010).
32. Weininger, D. SMILES a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–36 (1988).
33. Landrum, G. *RDKit: Open-source cheminformatics.* http://www.rdkit.org.
34. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17,** 490–519 (1996).
35. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminformatics* **6,** 1–4 (2014).
36. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3,** 33 (2011).
37. gnu.org. at http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html.
38. The BSD 3-Clause License | Open Source Initiative. http://opensource.org/licenses/BSD-3-Clause.
39. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1,** 140022 (2014).
40. Pettifer, S. R., Attwood, P. T. K. in *Encyclopedia of Systems Biology* (eds Dubitzky W., Wolkenhauer O., Cho K.-H. & Yokota H.) 1016–1016 (Springer: New York, 2013) http://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_1375.
41. Shao, Y. *et al.* Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113,** 184–215 (2015).
42. World Community Grid - http://www.worldcommunitygrid.org/.

Data Citations

1. Aspuru-Guzik, A. *figshare* https://dx.doi.org/10.6084/m9.figshare.1610063.v4 (2016).

Acknowledgements

A.A.-G. and E.O.P.-K. acknowledges the Department of Energy through grant DE-SC0008733 for funding. S.A.L. acknowledges the ORISE EERE Postdoctoral Fellowship for funding and support. K.L. and T.L. acknowledge Harvard College Research Project for financial aid. L.R.S. acknowledges the Harvard University Center for the Environment for a Research Fellowship. This research would not have been possible without the use of the Harvard FAS Odyssey Cluster and support from FAS Research Computing.

Author Contributions

S.L. contributed to the writing manuscript, analysed the data, computed all of the supplementary MP2 energies, and optimizations utilizing the $CPCM^{H2O}$ solvation model. E.P.K. analysed the data, aided in collecting literature data, contributed to running quantum-chemical calculations and wrote the Data Descriptor. G.N.S. aided in collecting literature data and contributed to running quantum-chemical calculations. K.L., T.L., L.R.S. and J.H. aided in collecting literature data. A.A.G. devised and supervised the project and aided in writing the Data Descriptor.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing financial interests: The authors declare no competing financial interests.

How to cite: Lopez, S. A. *et al.* The Harvard organic photovoltaic dataset *Sci. Data* 3:160086 doi: 10.1038/sdata.2016.86 (2016).