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

July 1, 2011

1 Molecular Library Generation

In Scheme S1, we show the set of building blocks. These are used to generate the molecular library. Further details of the molecular library generation will be detailed in a separate publication. [1]

2 Descriptors

We selected a subset of descriptors available through ChemAxon. The subset selected is shown in Table S1. These descriptors were coded in a program and applied to our huge set of molecules.

It should be noted that many descriptors are calculated for each atom of the molecule. In these cases we selected the lowest, the highest and the average value to be used in the calibration and in the prediction.

3 Calibration molecules

We selected a set of 50 molecules which have been studied experimentally and results are available.

We built molecular formulae from the original papers. For these calculations, aliphatic chains were stripped from the molecular formula. The set of molecules with efficiency parameters measured experimentally, thus used in this study, is shown in Fig. S1 and Fig. S2.

4 Selection of descriptors and results for the calibration molecules

Fig. S3, Fig. S4 and Fig. S5 show plots of V_{oc} versus each one of the 33 descriptors, none of them showing an obvious correlation. We used `R` to perform a multiple linear regression. As could have been expected, 50 experimental data with 33 descriptors make for a very good fit with R-squared of 0.9580. Some of the descriptors are not statistically significant, so we removed the less significant descriptors and ended up with

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

Table S1 The set of descriptors calculated using ChemAxon.

1. Molecular mass
2. logP partition constant octanol/water
3. Ring count
4. Hydrogen bond acceptor atom count in molecule
5. Hydrogen bond donor atom count in molecule
6. Rotatable bond count
7. Molecular polarizability
8. Refractivity
9. van der Waals surface area
10. van der Waals volume
11. Water accesible area calculation (ASA)
12. Water accesible area calculation of all atoms with positive partial charge (ASA+)
13. Water accesible area calculation of all atoms with negative partial charge (ASA-)
14. Water accesible area calculation of all hydrophobic ($\|q_i\| < 0.125$) atoms (ASA_H)
15. Water accesible area calculation of all polar ($\|q_i\| >= 0.125$) atoms (ASA_P)
16. Electrophilic localization energy (lowest)
17. Electrophilic localization energy (highest)
18. Electrophilic localization energy (average)
19. Partial charge calculation (lowest)
20. Partial charge calculation (highest)
21. Partial charge calculation (average)
22. Electron density (lowest)
23. Electron density (highest)
24. Electron density (average)
25. Steric hindrance (lowest)
26. Steric hindrance (highest)
27. Steric hindrance (average)
28. Orbital electronegativity (sigma) (lowest)
29. Orbital electronegativity (sigma) (highest)
30. Orbital electronegativity (sigma) (average)
31. Orbital electronegativity (pi) (lowest)
32. Orbital electronegativity (pi) (highest)
33. Orbital electronegativity (pi) (average)

a fitting from 20 descriptors (with R-squared of 0.9455) all of them significant. Table S2 show results of the fitting.

Next, we used the same technique for J_{sc}. Table S3 show results of the fitting. In this case, 18 significant descriptors fit the experimental values with R-squared of 0.8989.

Table S4 show results of the fitting of PCE. In this case, 15 significant descriptors fit the experimental values with R-squared of 0.8409.

Table S5 show results of the fitting of FF. In this case, 20 significant descriptors fit the experimental values with R-squared of 0.6170; not all of them are significant but one has to go to only 12 descriptors to get them all significant and then R-squared drops to 0.4757. This is interpreted as meaning that the fitting for FF is not credible at all, as it may be expected for this morphology based experimental parameter.

Table S6 show results of the fitting of the product $V_{oc}J_{sc}$. The fitting is good with R-squared of 0.8809 and all 20 descriptors are significant.

A common test of the calibration of the training set to the descriptors is the "leave 1/3 out" technique. The whole set of calibration molecules is randomly divided in three subsets: A (17 molecules in this case), B (17 molecules), and C (16 molecules). For V_{oc}, the set of 20 descriptors previously selected is used to fit

Table S2 Results for the fitting of V_{oc} to 20 descriptors

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	17.0788482	1.9956354	8.558	1.99e-09	***
logP	−0.1328980	0.0155249	−8.560	1.98e-09	***
RingCount	0.2204759	0.0279730	7.882	1.08e-08	***
AcceptorCount	−0.0996894	0.0139580	−7.142	7.35e-08	***
RotBondCount	0.2374734	0.0240348	9.880	8.67e-11	***
Refractivity	−0.0075611	0.0017250	−4.383	0.00014	***
VdWArea	0.0063744	0.0009059	7.036	9.72e-08	***
VdWVolume	−0.0083368	0.0013763	−6.057	1.36e-06	***
ASA	−0.0037351	0.0010949	−3.411	0.00192	**
ASAH	0.0024702	0.0010328	2.392	0.02347	*
ElecLocEn(lo)	0.0371339	0.0157293	2.361	0.02517	*
PartialCh(lo)	−1.6637886	0.4416521	−3.767	0.00075	***
PartialCh(hi)	−3.3145356	0.5805118	−5.710	3.54e-06	***
PartialCh(avg)	50.3586495	6.6593831	7.562	2.46e-08	***
ElecDen(lo)	−0.8402735	0.2704179	−3.107	0.00420	**
ElecDen(avg)	−2.3826616	0.2394278	−9.951	7.37e-11	***
StericHind(hi)	−0.8950969	0.1545381	−5.792	2.82e-06	***
σElecNeg(lo)	0.1993522	0.0588838	3.386	0.00206	**
σElecNeg(hi)	0.0448449	0.0190354	2.356	0.02545	*
σElecNeg(avg)	−1.4447603	0.2073776	−6.967	1.17e-07	***
πElecNeg(hi)	0.2316545	0.0345617	6.703	2.36e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04702 on 29 degrees of freedom
(8 observations deleted due to missingness)
Multiple R-squared: 0.9455, Adjusted R-squared: 0.9079
F-statistic: 25.15 on 20 and 29 DF, p-value: 4.09e-13

separately to the subsets AB (34 molecules), AC (33 molecules) and BC (33 molecules). The coefficients thus obtained are used to predict J_{sc} for the set not used in the fitting (C for AB, for instance). Results are shown in Fig. S6. As can be seen, the prediction of excluded molecules is consistent.

We performed the same analysis for J_{sc} —results are shown in Fig. S7— and for the product $V_{oc}J_{sc}$ —results shown in Fig. S8.

Table S3 Results for the fitting of J_{sc} to 18 descriptors

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	241.2144	34.5775	6.976	7.91e-08	***
logP	1.5085	0.4307	3.502	0.001424	**
AcceptorCount	1.6288	0.4314	3.776	0.000678	***
DonorCount	6.0441	1.6055	3.765	0.000700	***
RotBondCount	2.3886	0.3408	7.008	7.24e-08	***
MolPolarizability	−0.3654	0.1033	−3.537	0.001297	**
ASA+	−174.9657	43.9844	−3.978	0.000388	***
ASA−	−174.9481	43.9840	−3.978	0.000389	***
ASAH	174.9342	43.9859	3.977	0.000389	***
ASAP	175.0909	43.9871	3.981	0.000386	***
ElecLocEn(lo)	−24.2310	3.4067	−7.113	5.42e-08	***
ElecLocEn(avg)	15.4843	2.4773	6.250	6.03e-07	***
ElecDen(lo)	−24.1885	6.1314	−3.945	0.000426	***
StericHind(hi)	16.6785	2.8987	5.754	2.48e-06	***
σElecNeg(lo)	−6.2720	1.4874	−4.217	0.000199	***
σElecNeg(hi)	1.1994	0.5035	2.382	0.023531	*
σElecNeg(avg)	−38.4895	5.5345	−6.954	8.40e-08	***
πElecNeg(hi)	2.5199	0.9816	2.567	0.015305	*
πElecNeg(avg)	23.4035	3.7781	6.195	7.06e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.356 on 31 degrees of freedom
(8 observations deleted due to missingness)
Multiple R-squared: 0.8989, Adjusted R-squared: 0.8402
F-statistic: 15.31 on 18 and 31 DF, p-value: 1.016e-10

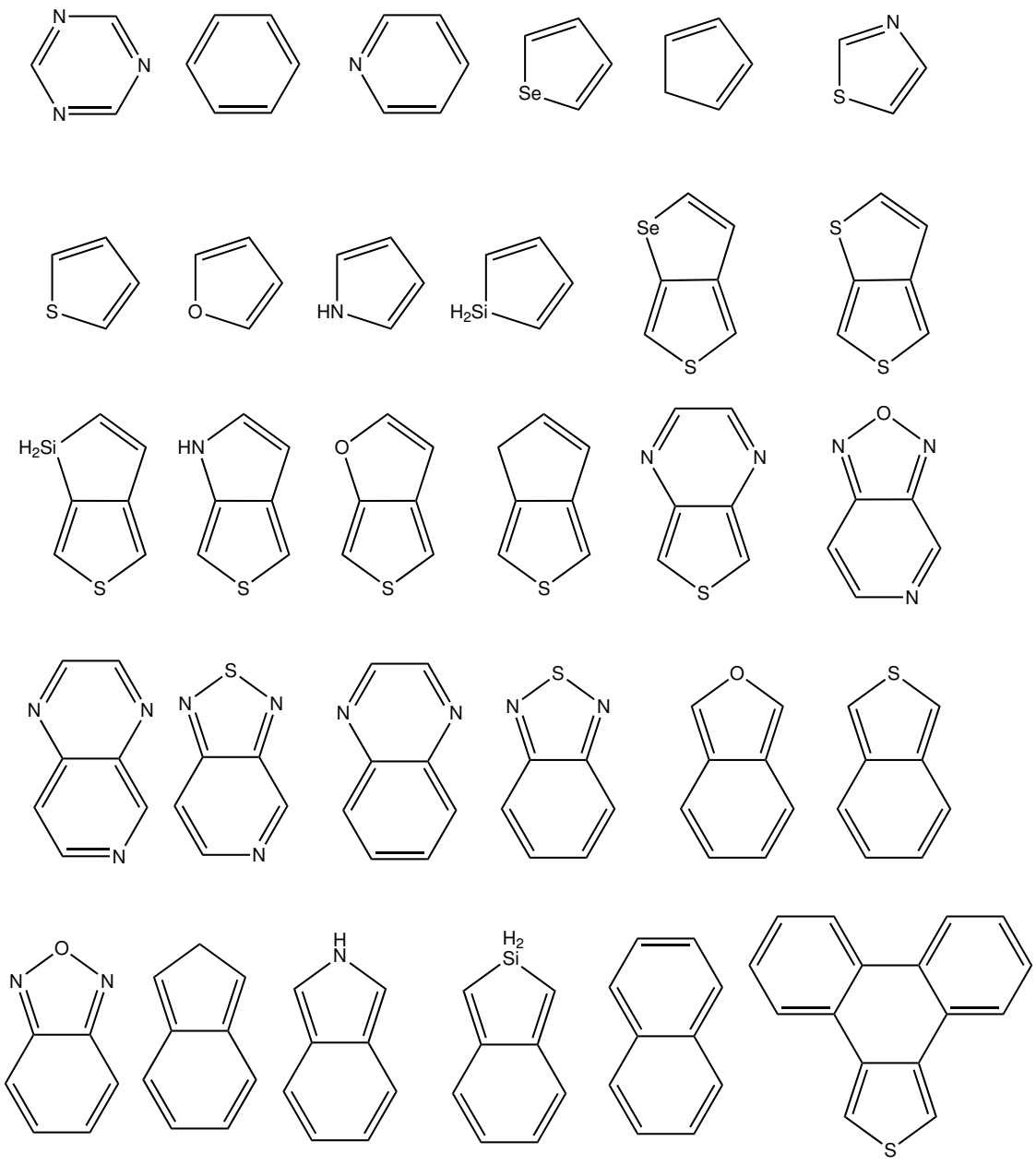
References

[1] *In Preparation*, 2011.

Table S4 Results for the fitting of %PCE to 15 descriptors

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	121.834817	22.037026	5.529	3.87e-06	***
RotBondCount	0.839267	0.162001	5.181	1.08e-05	***
Refractivity	−0.043604	0.014191	−3.073	0.004234	**
VdWArea	0.019545	0.006214	3.145	0.003503	**
ASA+	−89.561436	20.122423	−4.451	9.21e-05	***
ASA−	−89.554274	20.122510	−4.450	9.22e-05	***
ASAH	89.537003	20.122831	4.450	9.24e-05	***
ASAP	89.592594	20.123894	4.452	9.18e-05	***
ElecLocEn(hi)	0.570364	0.165122	3.454	0.001535	**
PartCharge(avg)	164.421493	82.195338	2.000	0.053744	.
ElecDen(lo)	−11.329732	2.569477	−4.409	0.000104	***
StericHind(lo)	−7.596139	2.418105	−3.141	0.003539	**
σElecNeg(hi)	0.523366	0.216473	2.418	0.021304	*
σElecNeg(avg)	−15.565577	1.935828	−8.041	2.81e-09	***
πElecNeg(hi)	1.782254	0.462344	3.855	0.000507	***
πElecNeg(avg)	4.212309	1.925005	2.188	0.035839	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.6436 on 33 degrees of freedom
(9 observations deleted due to missingness)
Multiple R-squared: 0.8409, Adjusted R-squared: 0.7686
F-statistic: 11.63 on 15 and 33 DF, p-value: 3.784e-09



Scheme S1 The thirty basic heterocyclic units used in this study are shown with their Mg chemical handles at the 2,5- or para- position of the five-membered and six-membered ring, respectively.

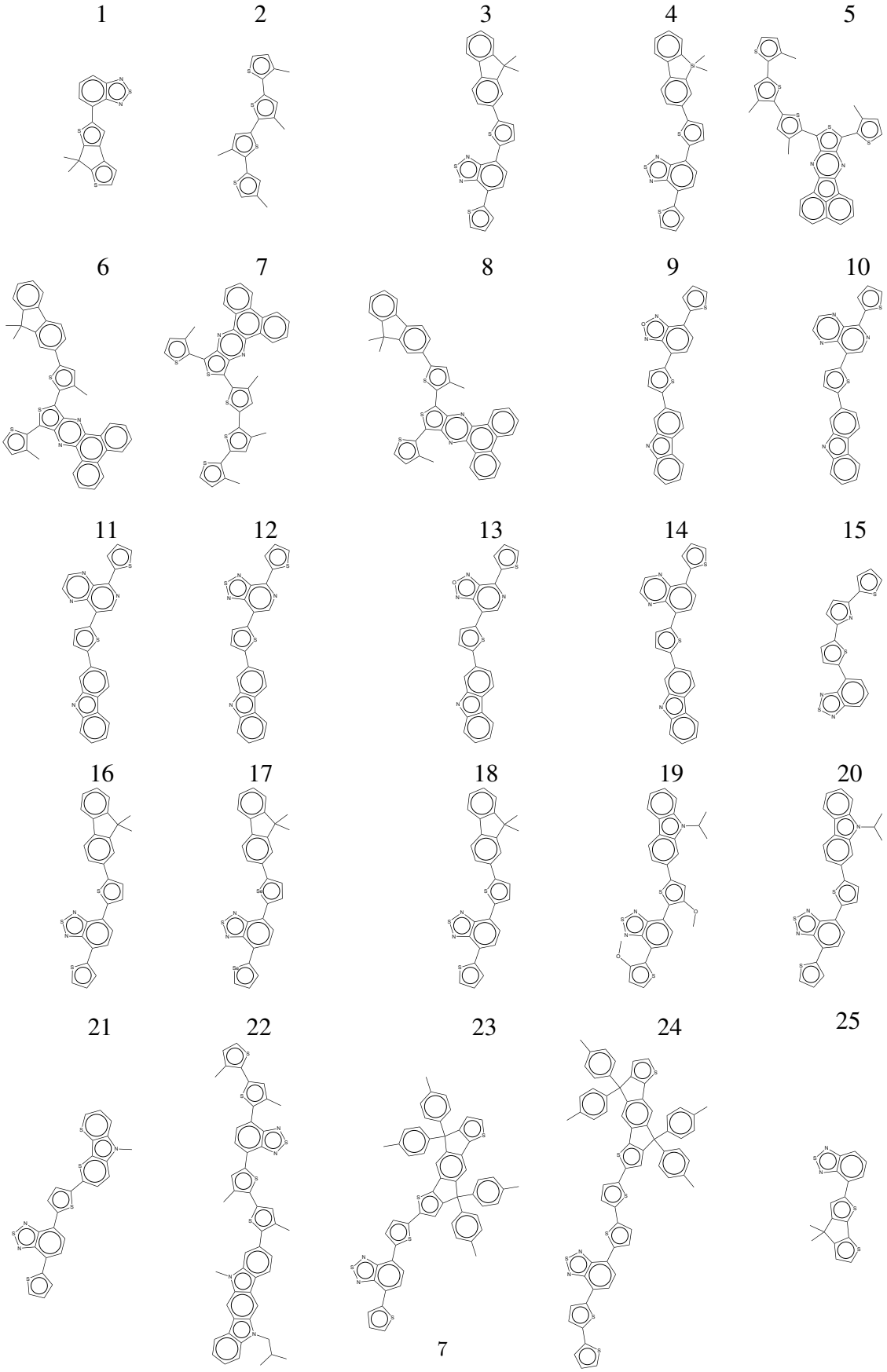


Fig. S1 First set of the fifty experimentally characterized molecules (shown here without aliphatic side chains). Once stripped of the aliphatic tails, several molecules are equivalent (6 and 8; 3, 16 and 18; and 48 and 50), but all are retained in the training set.

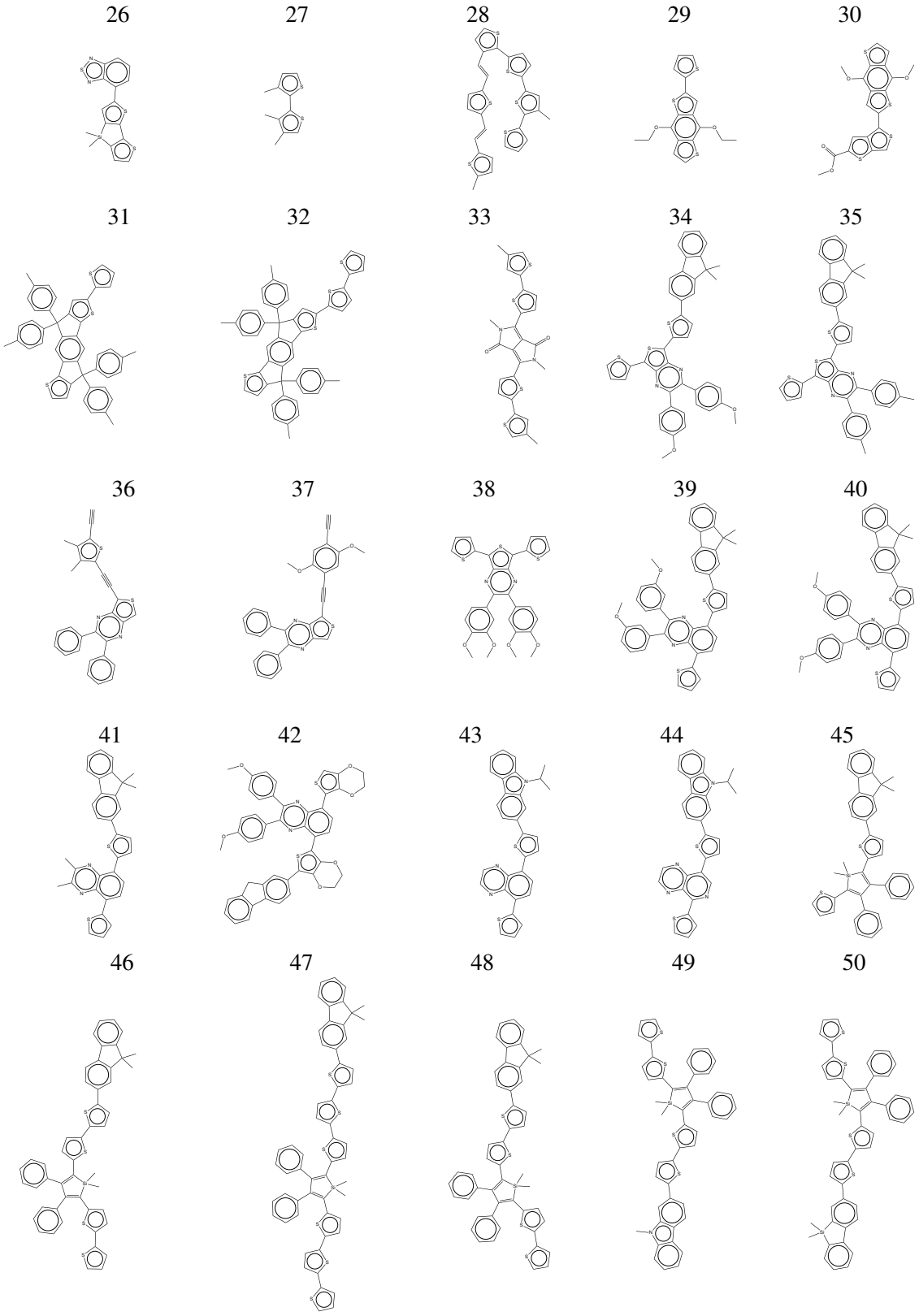


Fig. S2 Second set of the fifty experimentally characterized molecules (shown here without aliphatic side chains). Once stripped of the aliphatic tails, several molecules are equivalent (6 and 8; 3, 16 and 18; and 48 and 50), but all are retained in the training set.

Table S5 Results for the fitting of FF to 20 descriptors. Note that not all of them are significant.

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	0.179217	0.937496	0.191	0.84978	
MolecMass	0.003691	0.001849	1.997	0.05568	.
RingCount	−0.335110	0.105620	−3.173	0.00365	**
DonorCount	−0.260610	0.097487	−2.673	0.01239	*
RotBondCount	−0.107250	0.038108	−2.814	0.00884	**
MolPolarizability	0.058696	0.018165	3.231	0.00315	**
VdWVolume	−0.005359	0.003037	−1.765	0.08854	.
ASA	7.987770	3.732320	2.140	0.04119	*
ASA+	−7.988750	3.732284	−2.140	0.04117	*
ASA-	−7.988540	3.732144	−2.140	0.04117	*
ASAP	0.001865	0.001394	1.338	0.19150	
ElecLocEn(lo)	0.478993	0.161497	2.966	0.00611	**
ElecLocEn(hi)	0.234037	0.068207	3.431	0.00188	**
ElecLocEn(avg)	−0.582206	0.201671	−2.887	0.00742	**
PartialCharge(lo)	−1.450116	1.032790	−1.404	0.17129	
PartialCharge(avg)	−14.432399	8.624697	−1.673	0.10539	
ElecDen(hi)	−1.170898	0.404904	−2.892	0.00733	**
ElecDen(avg)	1.407407	0.386459	3.642	0.00109	**
σElecNeg(hi)	0.077231	0.030475	2.534	0.01715	*
πElecNeg(lo)	−0.046796	0.041594	−1.125	0.27011	
πElecNeg(hi)	−0.133699	0.070364	−1.900	0.06776	.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 0.07763 on 28 degrees of freedom					
(9 observations deleted due to missingness)					
Multiple R-squared: 0.6170, Adjusted R-squared: 0.3434					
F-statistic: 2.255 on 20 and 28 DF, p-value: 0.02361					

Table S6 Results for the fitting of $V_{oc}J_{sc}$to 20 descriptors

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	171.92666	30.65770	5.608	4.69e-06	***
MolMass	0.03479	0.01458	2.386	0.023758	*
logP	0.96304	0.28544	3.374	0.002120	**
AcceptorCount	1.11779	0.31584	3.539	0.001375	**
DonorCount	3.49484	1.37146	2.548	0.016383	*
RotBondCount	1.94840	0.28063	6.943	1.25e-07	***
Refractivity	-0.16140	0.03961	-4.075	0.000327	***
ASA	134.93698	47.32570	2.851	0.007942	**
ASA+	-134.97116	47.32624	-2.852	0.007929	**
ASA-	-134.95525	47.32343	-2.852	0.007932	**
ASAP	0.07088	0.02203	3.218	0.003167	**
ElecLocEn(lo)	-12.30036	2.60418	-4.723	5.47e-05	***
ElecLocEn(avg)	11.63824	2.18073	5.337	9.95e-06	***
PartCharge(hi)	-46.00685	12.68492	-3.627	0.001090	**
ElecDen(lo)	-20.96171	5.18290	-4.044	0.000355	***
ElecDen(avg)	-9.95094	5.09448	-1.953	0.060495	.
StericHind(hi)	11.98562	2.77777	4.315	0.000169	***
σElecNeg(lo)	-5.57994	1.35516	-4.118	0.000290	***
σElecNeg(avg)	-23.42843	3.95824	-5.919	1.99e-06	***
πElecNeg(hi)	3.08369	0.75953	4.060	0.000340	***
πElecNeg(avg)	13.62962	2.72602	5.000	2.54e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.042 on 29 degrees of freedom
(8 observations deleted due to missingness)
Multiple R-squared: 0.8809, Adjusted R-squared: 0.7988
F-statistic: 10.73 on 20 and 29 DF, p-value: 1.427e-08

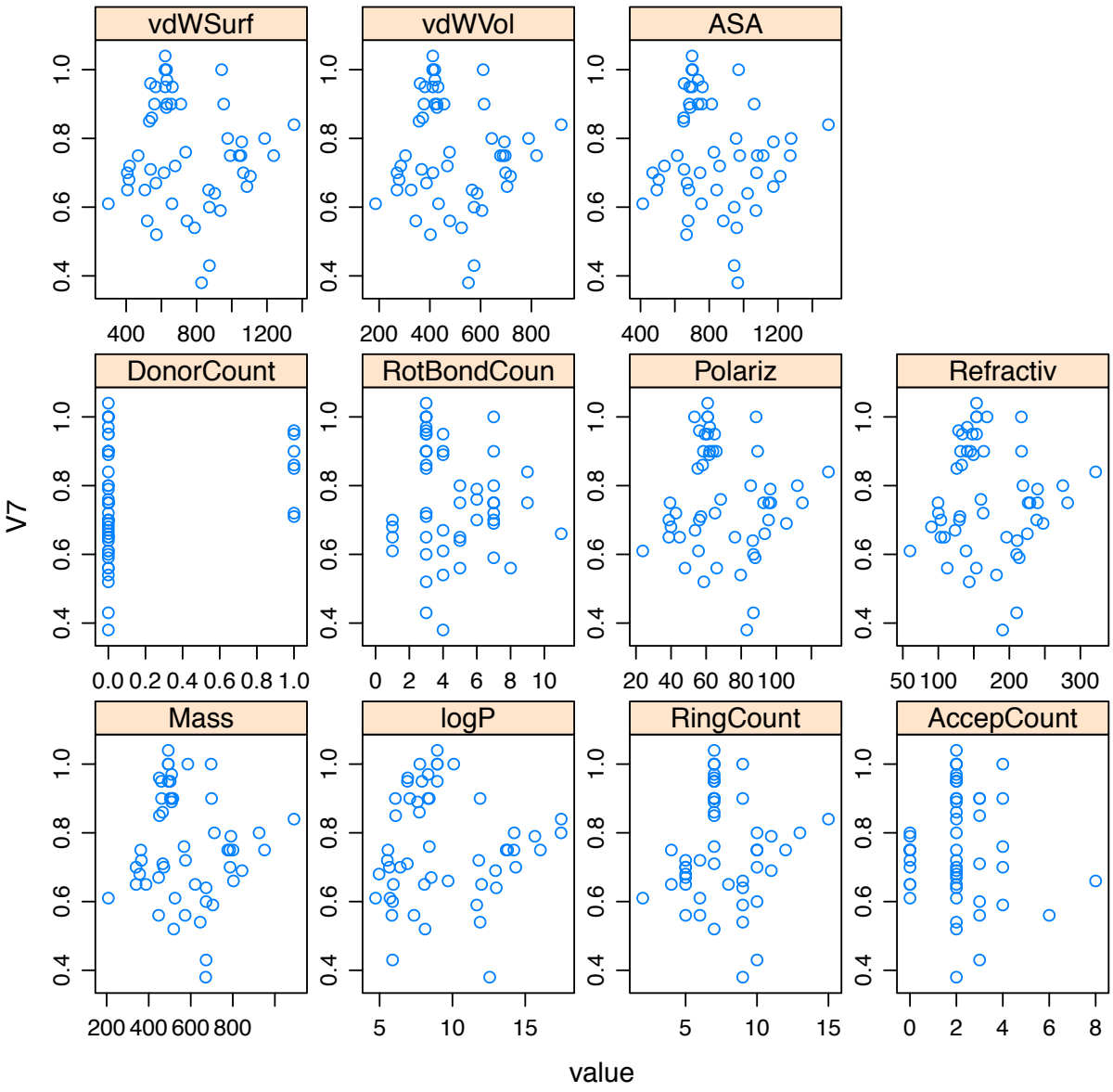


Fig. S3 Experimental values of V_{oc} versus a set of 33 descriptors obtained from ChemAxon for the set of 50 calibration molecules stripped of their aliphatic tails.

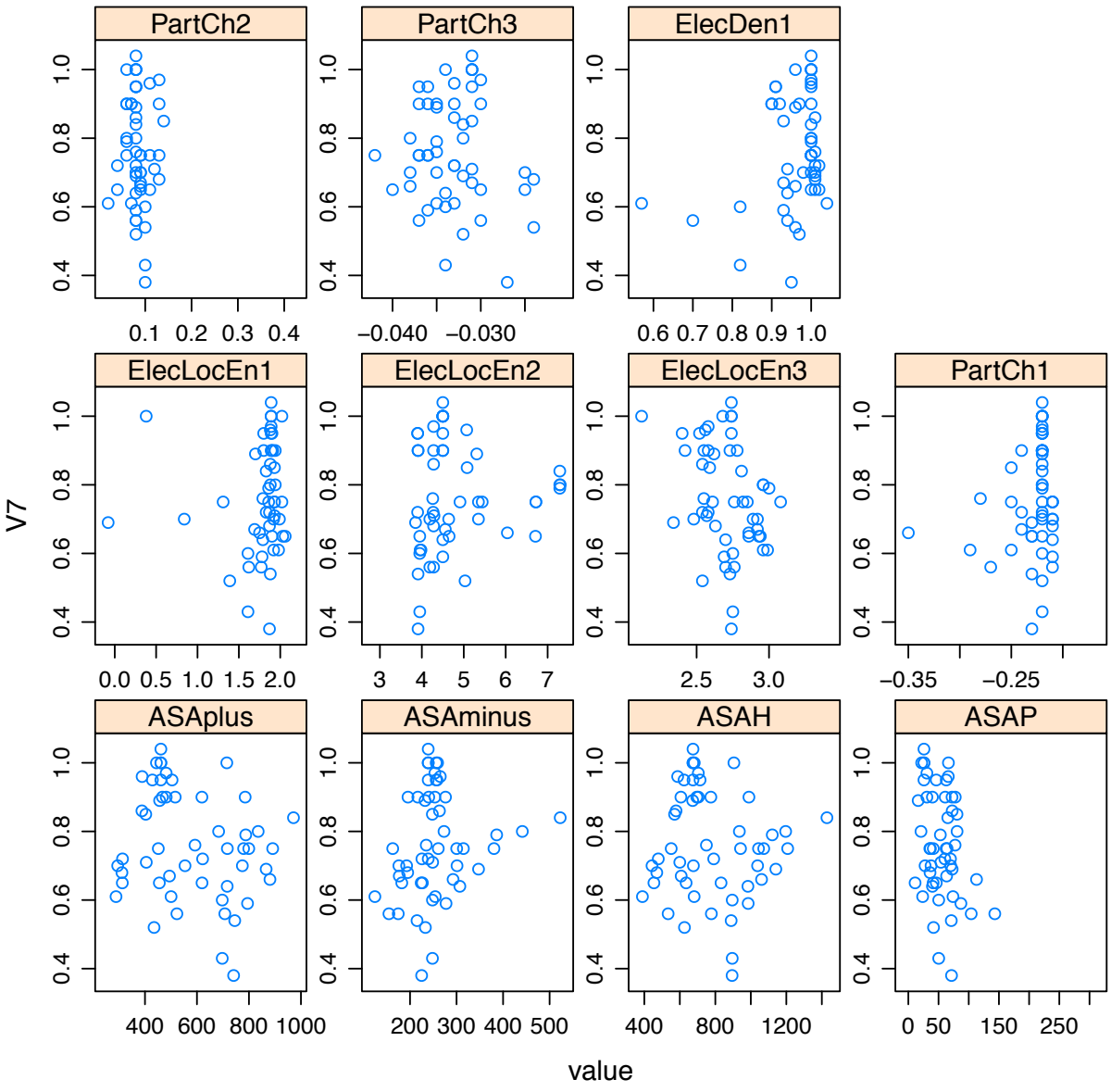


Fig. S4 Experimental values of V_{oc} versus a set of 33 descriptors obtained from ChemAxon for the set of 50 calibration molecules stripped of their aliphatic tails.

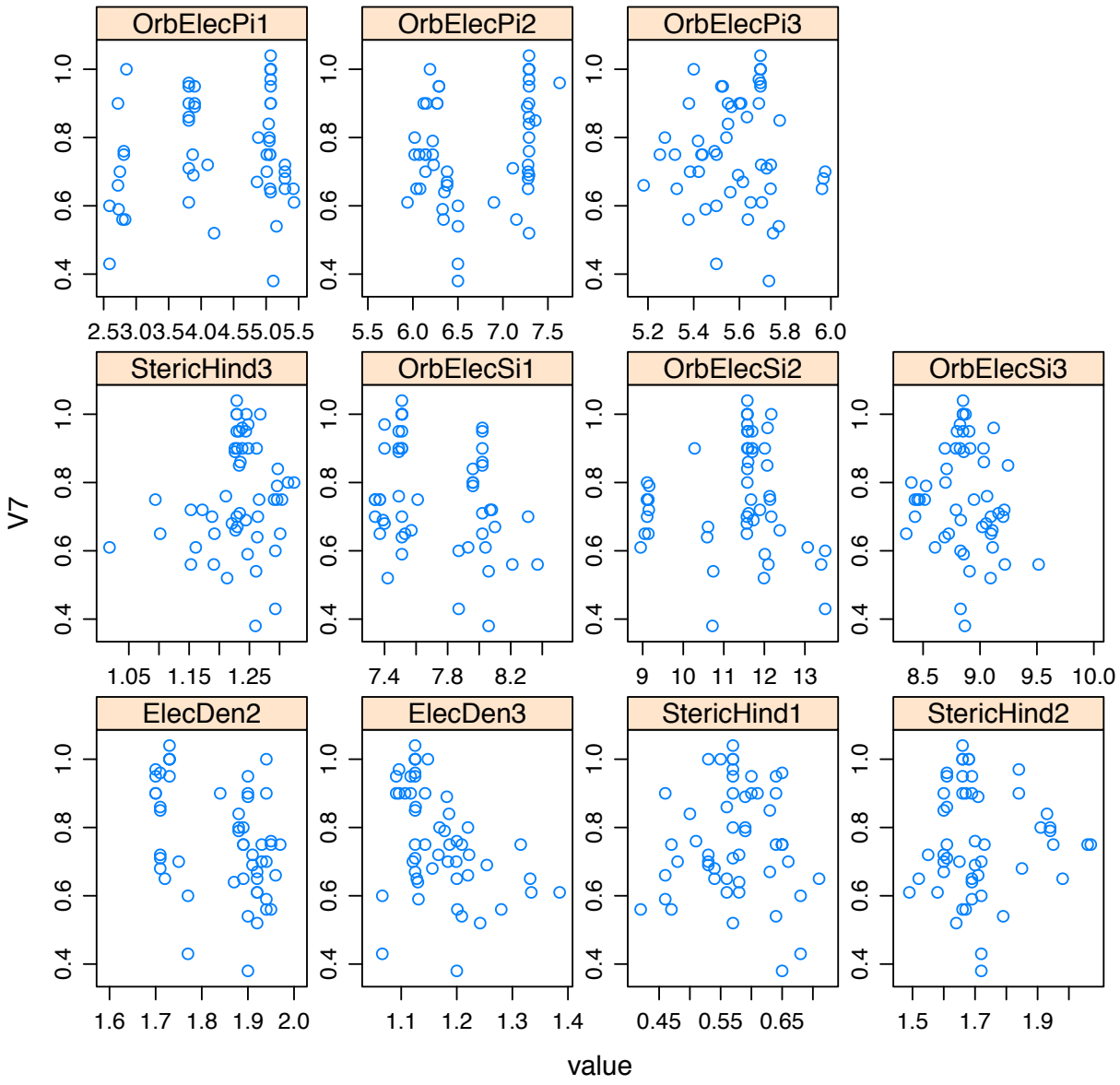


Fig. S5 Experimental values of V_{oc} versus a set of 33 descriptors obtained from ChemAxon for the set of 50 calibration molecules stripped of their aliphatic tails.

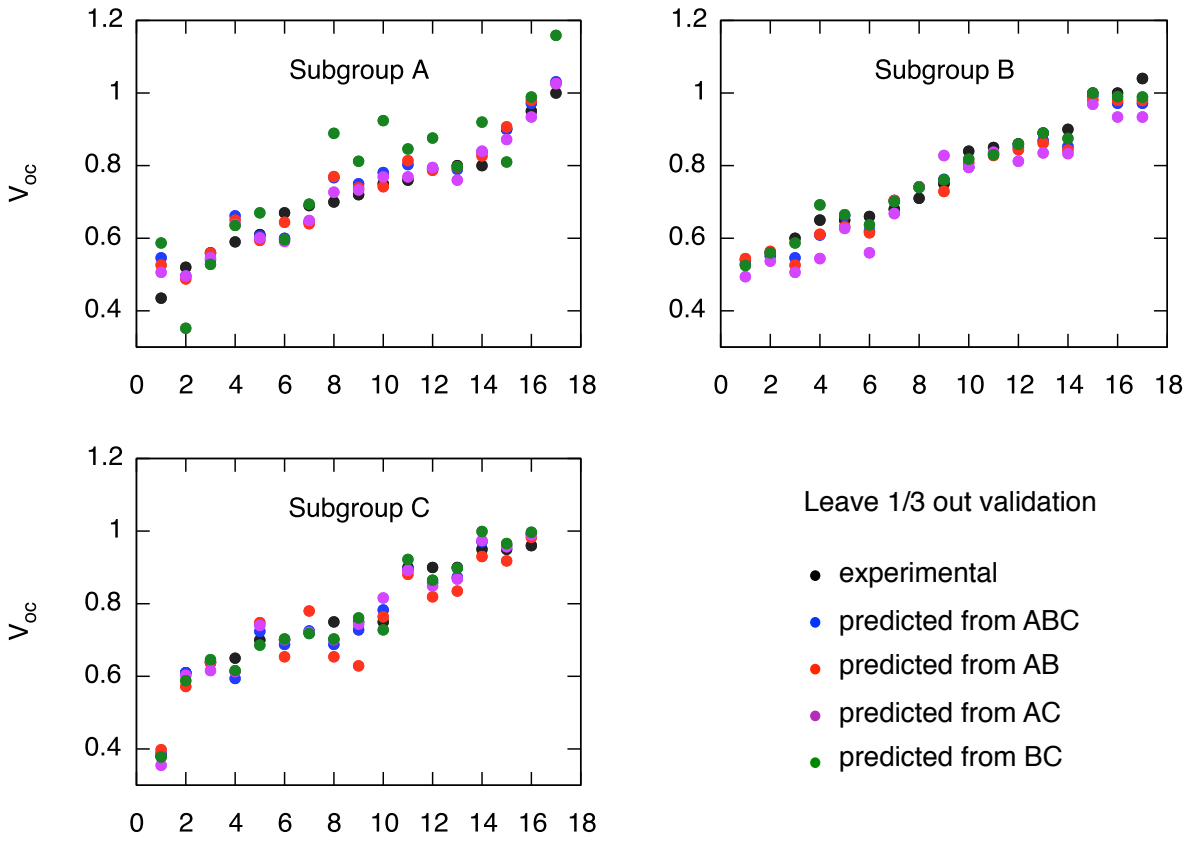


Fig. S6 Results of the "leave 1/3 out" correlation test for V_{oc}.

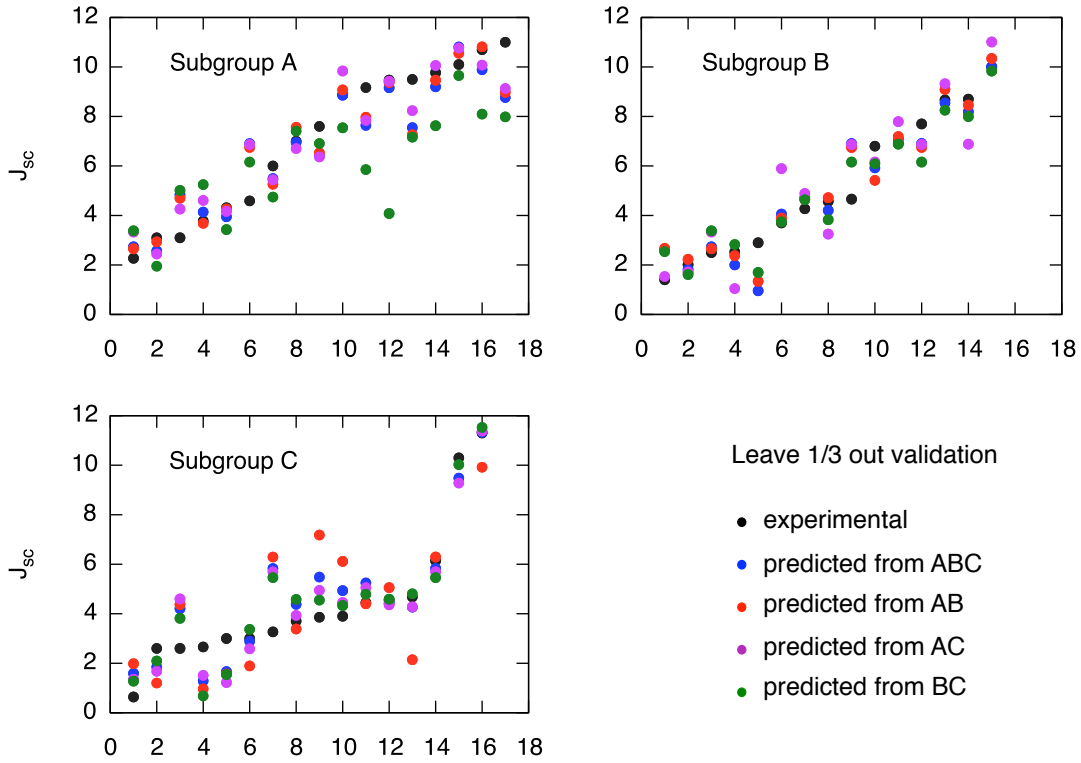


Fig. S7 Results of the "leave 1/3 out" correlation test for J_{sc}.

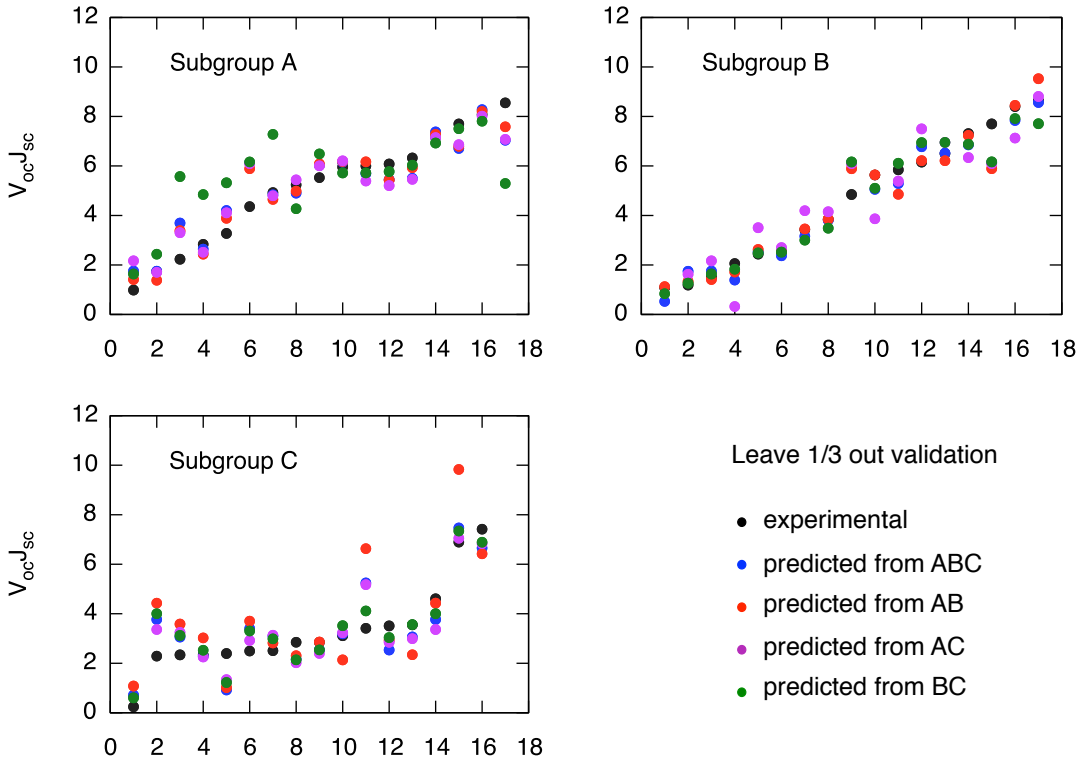


Fig. S8 Results of the "leave 1/3 out" correlation test for $V_{\mathrm{oc}}J_{\mathrm{sc}}$.