Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.
This journal is © The Royal Society of Chemistry 2024

Supplementary Information:

Beyond Molecular Structure: Critically Assessing Machine Learning for Designing Organic Photovoltaic Materials and Devices

Martin Seifrid[a,b†*], Stanley Lo[b,†], Dylan G. Choi[c], Gary Tom[b,d,e], My Linh Le[f], Kunyu Li[c,‡], Rahul Sankar[c,‡], Hoai-Thanh Vuong[c,‡], Hiba Wakidi[c,‡], Ahra Yi[c,‡], Ziyue Zhu[c,‡], Nora Schopp[c], Aaron Peng[c], Benjamin R. Luginbuhl[c], Thuc-Quyen Nguyen[c*], Alán Aspuru-Guzik[b,d,e,g,h,i,j*]

[a]Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, USA
[b]Department of Chemistry, Chemical Physics Theory Group, 80 St. George St., University of Toronto, Ontario M5S 3H6, Canada
[c]Department of Chemistry and Biochemistry, Center for Polymers and Organic Solids, University of California Santa Barbara, Santa Barbara, California 93106, USA
[d]Department of Computer Science, University of Toronto, 40 St George St, Toronto, ON M5S 2E4
[e]Vector Institute for Artificial Intelligence, 661 University Ave. Suite 710, Toronto, Ontario M5G 1M1, Canada
[f]Materials Department, University of California Santa Barbara, Santa Barbara, California 93106, USA
[g]Department of Chemical Engineering & Applied Chemistry, 200 College St., University of Toronto, Ontario M5S 3E5, Canada
[h]Department of Materials Science & Engineering, 184 College St., University of Toronto, Ontario M5S 3E4, Canada
[i]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), 661 University Ave., Toronto, Ontario M5G 1M1, Canada
[j]Acceleration Consortium, University of Toronto, 80 St. George St, Toronto, ON M5S 3H6, Canada
[†,‡]Authors contributed equally

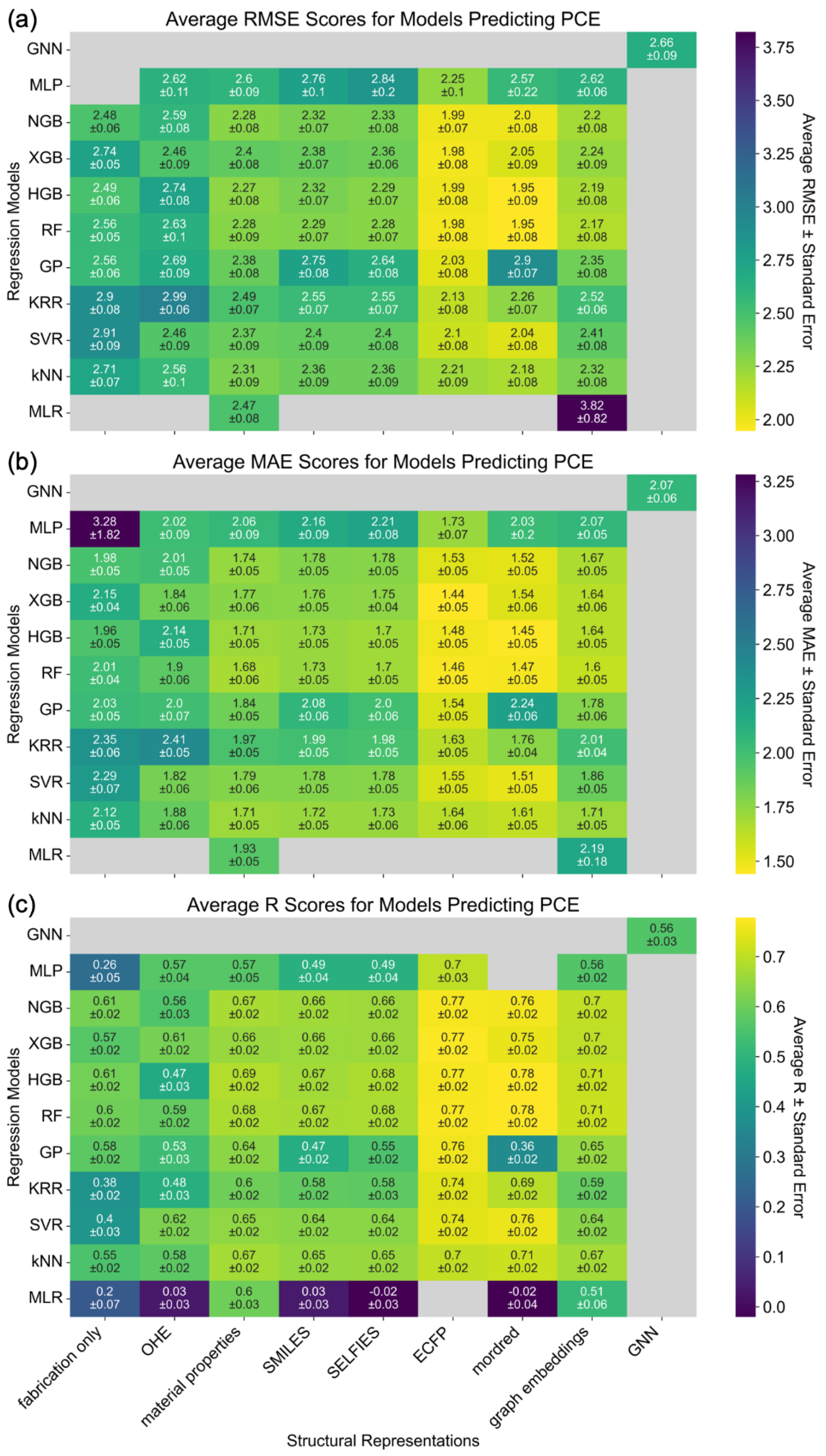

Figure S1. Heatmap of model performance for predicting PCE from the molecular structure of the donor and acceptor materials using various models and representations as measured by (a) RMSE, (b) MAE, and (c) R scores.

Table S1. Performance metrics of GNN models from which graph embeddings were extracted.

Target	Score	Donors	Acceptors
HOMO	R^2	-0.09	0.21
	MSE	1.24	0.46
LUMO	R^2	0.01	0.54
	MSE	0.70	0.51
E_g^{opt}	R^2	0.43	0.68
	MSE	0.49	0.31
Best loss		0.76	0.42

Table S2. Table of model performance for the PUFp representation.

Model	R2	RMSE	MAE	R
MLR				0.02±0.04
kNN	0.43±0.03	2.32±0.07	1.76±0.06	0.66±0.02
SVR	0.43±0.03	2.32±0.23	1.74±0.06	0.67±0.02
KRR	0.45±0.03	2.29±0.06	1.76±0.05	0.68±0.02
GP	0.35±0.03	2.47±0.08	1.87±0.06	0.61±0.02
RF	0.48±0.02	2.22±0.07	1.67±0.05	0.70±0.02
HGB	0.40±0.03	2.37±0.06	1.85±0.05	0.65±0.02
XGB	0.47±0.02	2.24±0.07	1.71±0.05	0.70±0.02
NGB	0.42±0.02	2.35±0.07	1.85±0.05	0.66±0.02
MLP	0.37±0.04	2.44±0.09	1.91±0.08	0.65±0.02

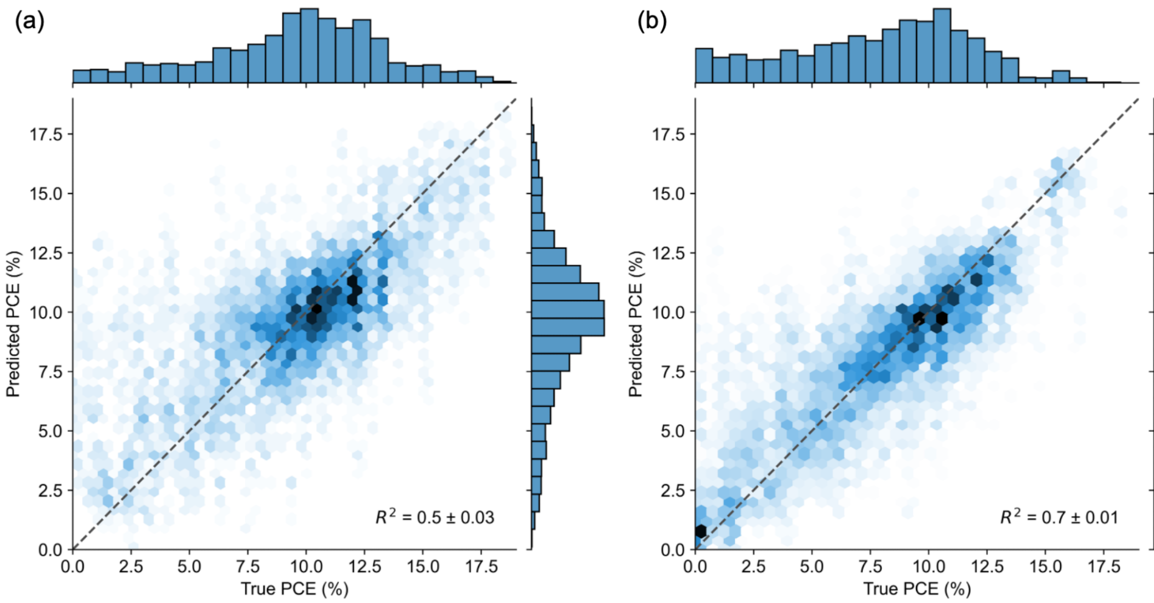

Figure S2. Heatmap parity plots of HGBR model predictions using 5-fold CV initialized from 7 random seeds. (a) The HGBR algorithm outperforms RF models trained on computed descriptors from the 2023 Greenstein and Hutchison dataset[1] (ca. 1000 unique D:A pairs) using much simpler molecular representations (ECFP and incomplete material properties). (b) The HGBR algorithm performs comparably to RF models trained on mordred descriptors and material properties from the 2021 Miyake and Saeki dataset[2] (ca. 1300 data points) using ECFP and material properties.

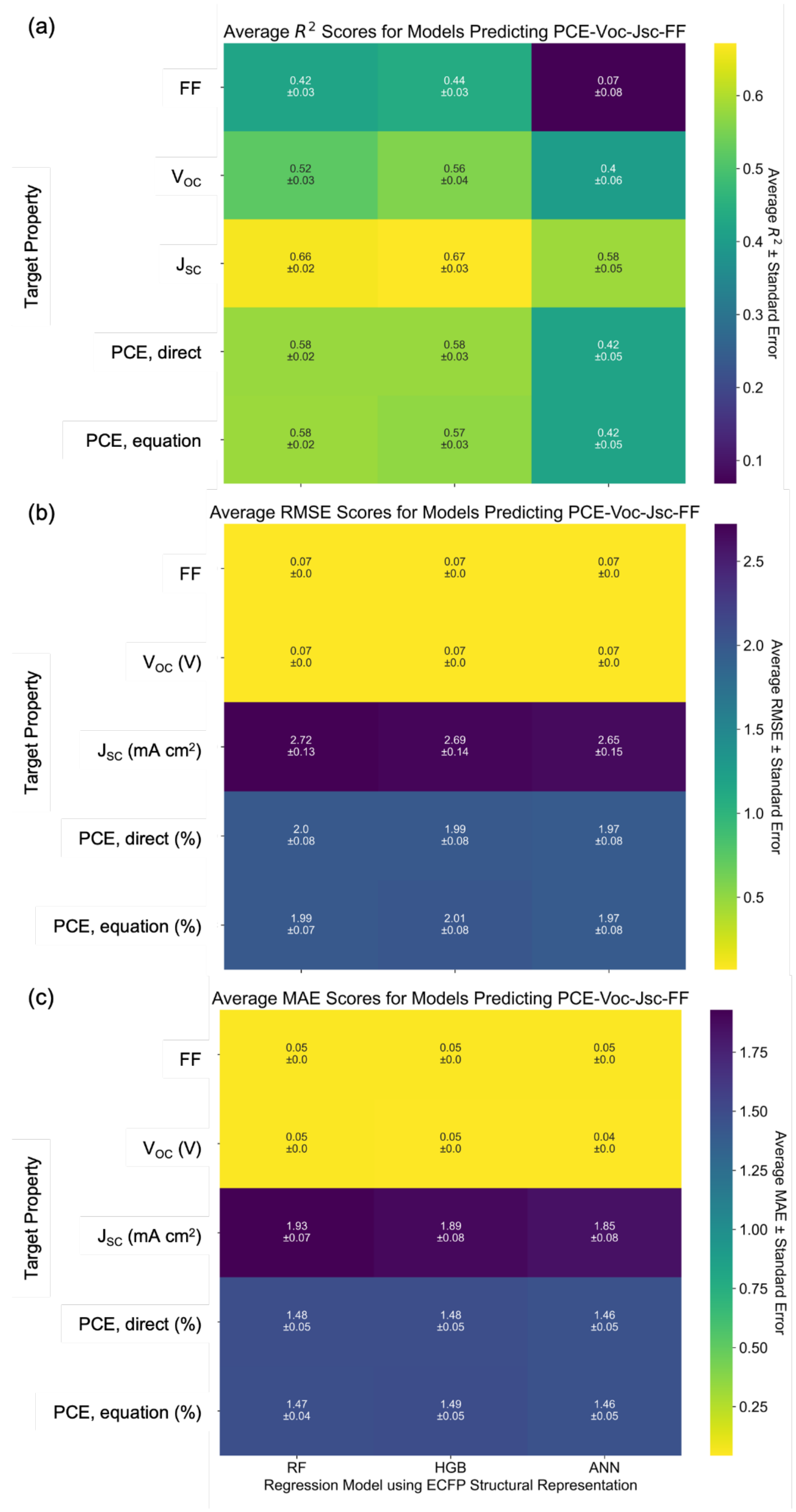

Figure S3. Heatmaps of performance of multi-output RF, HGB, and ANN models trained on the ECFP molecular representations for predicting FF, V_{OC}, J_{SC}, PCE (directly), and PCE (from Equation 1 in the main text) concurrently, as measure by (a) R^2, (b) RMSE, (c) MAE scores.

Table S3. Encodings of HTLs and ETLs along with their energy level, measurement method, and reference. CV: cyclic voltammetry; KP: Kelvin probe; UPS: ultraviolet photon spectroscopy; WF: work function.

Name	Energy (eV)	Energy level	Measurement technique	Reference
C_{60}-bissalt	-3.9	LUMO	CV	[3]
Ca	-2.9	WF		[4]
EDTA-ZnO	-4.05	WF	UPS	[5]
FPI	-3.93	LUMO	CV	[3]
LiF	-3	WF		
Al	-4.3	WF		[6]
N719:PrC_{60}MAI	-3.9	WF	UPS	[7]
PDIN	-3.6			
PDINO	-3.9			[6]
PEIE	-4.1	WF		[8]
PFN	-2.14	LUMO	CV	[9]
ZnO/PFN	-2.14	LUMO	CV	[9]
PFN-Br	-2.18	LUMO	CV	[10]
PFNDI-Br	-4.18	LUMO	CV	[11]
PNDIT-F3N	-3.9	LUMO	CV	[11]
PNDIT-F3N-Br	-4.18	LUMO	CV	[11]
TiO_2:TOPD	-4.23	WF	KP	[12]
ZnO	-4.46	WF		[13]
ZnO/PFN-Br	-4.08	LUMO	UPS + E_g^{opt}	[10]
ZrAcAc	-4.65	WF	UPS	[14]
CuSCN	-5.3	WF		[15]
MoOx	-5.3	WF	KP	[16]
PEDOT:PSS	-5.2	WF	UPS	[13]
V_2O_5	-4.7			[17]

Table S4. Physicochemical descriptors of solvents and solvent additives from HSPiP.[18]

Abbreviation	**Descriptor**	**Units**
dipole	Electric dipole moment	D
dD	Hansen solubility parameter: energy from dispersion forces	$MPa^{0.5}$
dP	Hansen solubility parameter: energy from intermolecular forces	$MPa^{0.5}$
dH	Hansen solubility parameter: energy from hydrogen bonds	$MPa^{0.5}$
dHDon	Hansen solubility parameter: hydrogen bonding donor term	$MPa^{0.5}$
dHAcc	Hansen solubility parameter: hydrogen bonding acceptor term	$MPa^{0.5}$
MW	Molecular weight	
Density	Density at 25 ºC	g/cm^3
BPt	Boiling point	ºC
MPt	Melting point	ºC
logKow	Log of octanol-water partition coefficient	
RI	Refractive index	
Trouton	Trouton's rule: (HVap at BPt) / BPt	J/kmol
RER	Relative evaporation rate	
ParachorGA	Parachor for surface tension	
RD	Molecular refractivity	
DCp	Heat capacity at 25 ºC	J/kmol
log n	Log of viscosity at 25 ºC	mPa•s
SurfTen	Surface tension at 25 ºC	mN/m

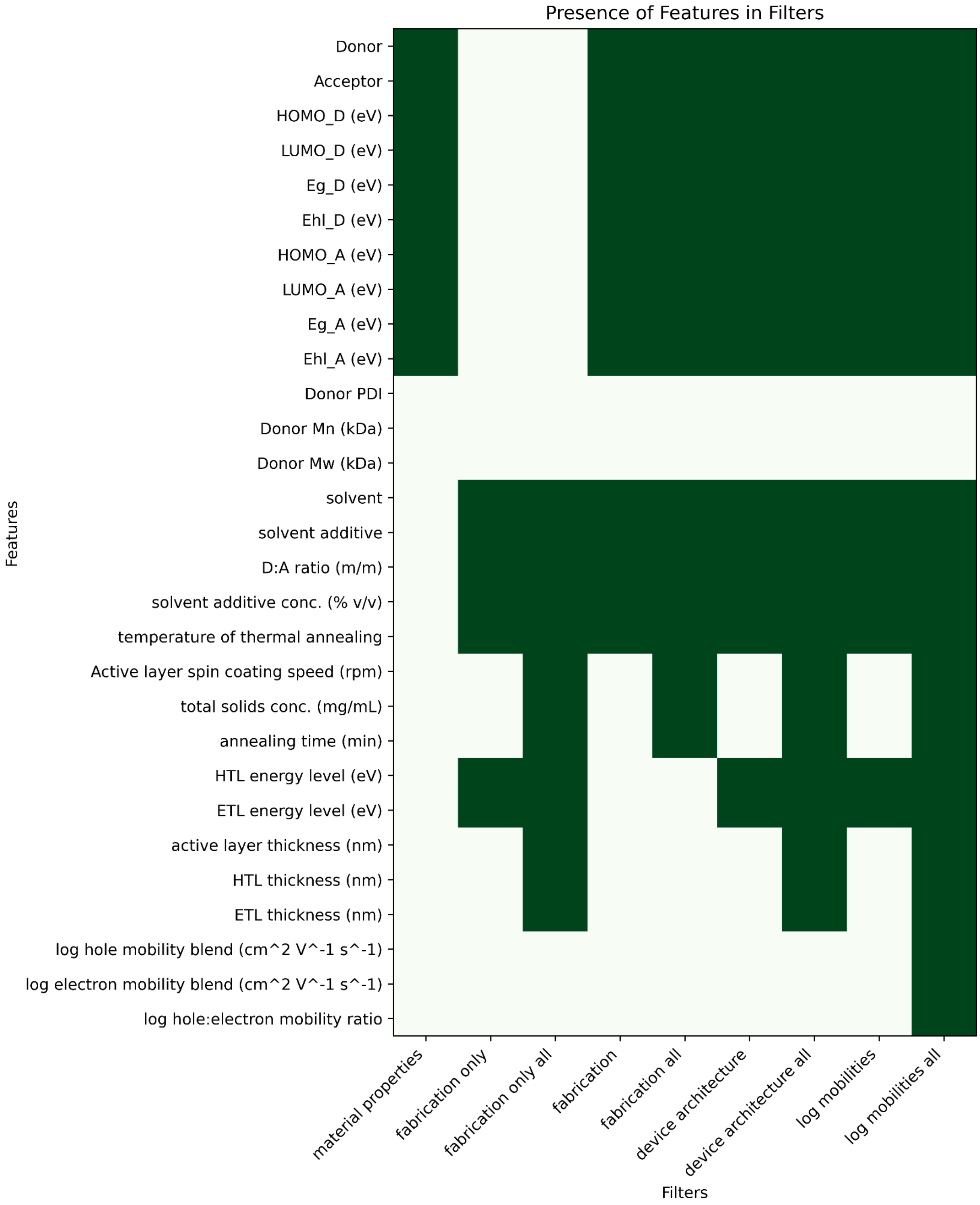

Figure S4. Plot of dataset features in the various filters.

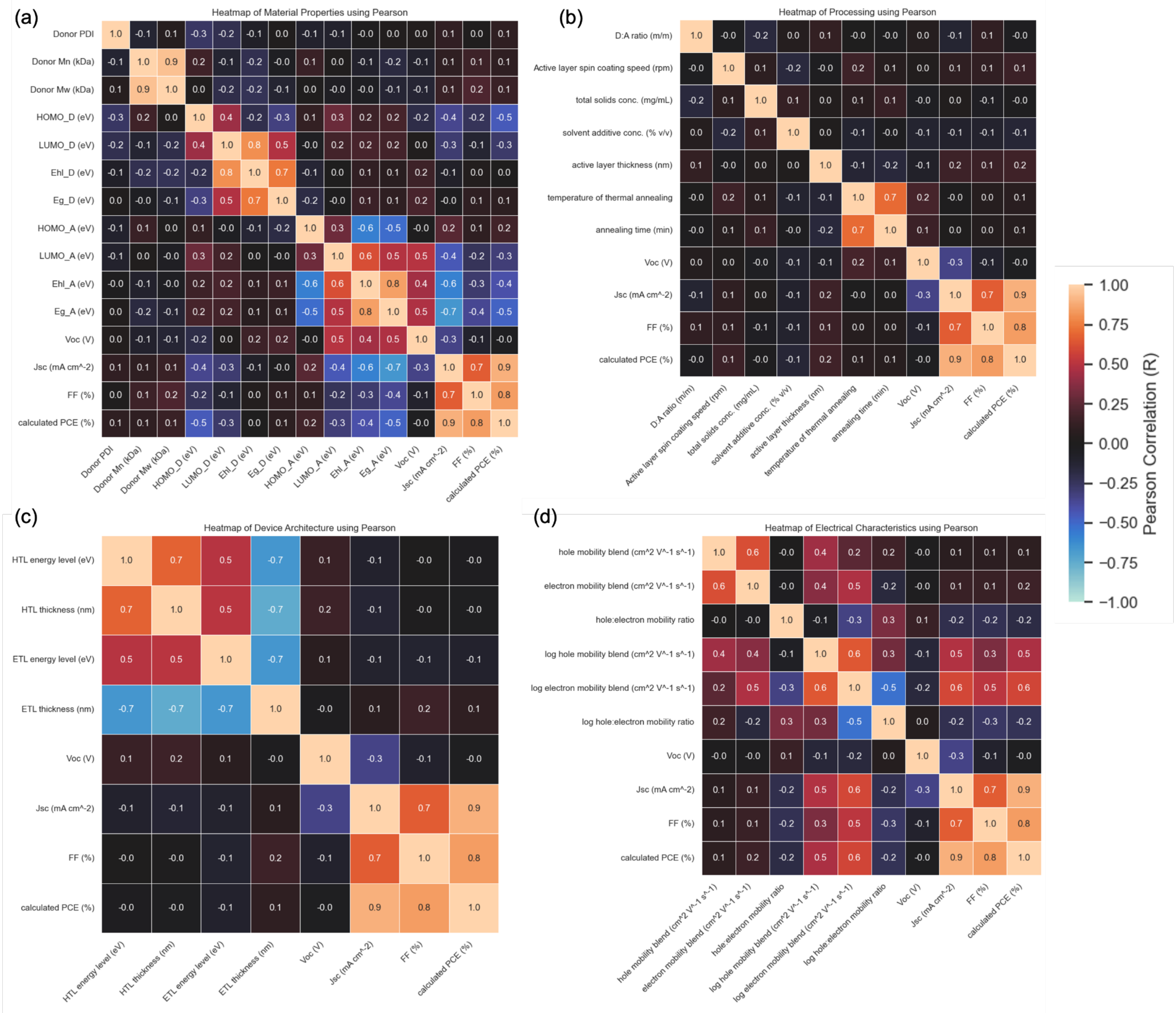

Figure S5. Pearson correlation coefficient heatmaps of the different groups of processing features with the output variables: (a) material properties, (b) processing, (c) device architecture, (d) charge carrier mobilities. Feature groups correspond to those defined in Figure S4. Correlation coefficient values in each cell are rounded to the nearest tenth for readability.

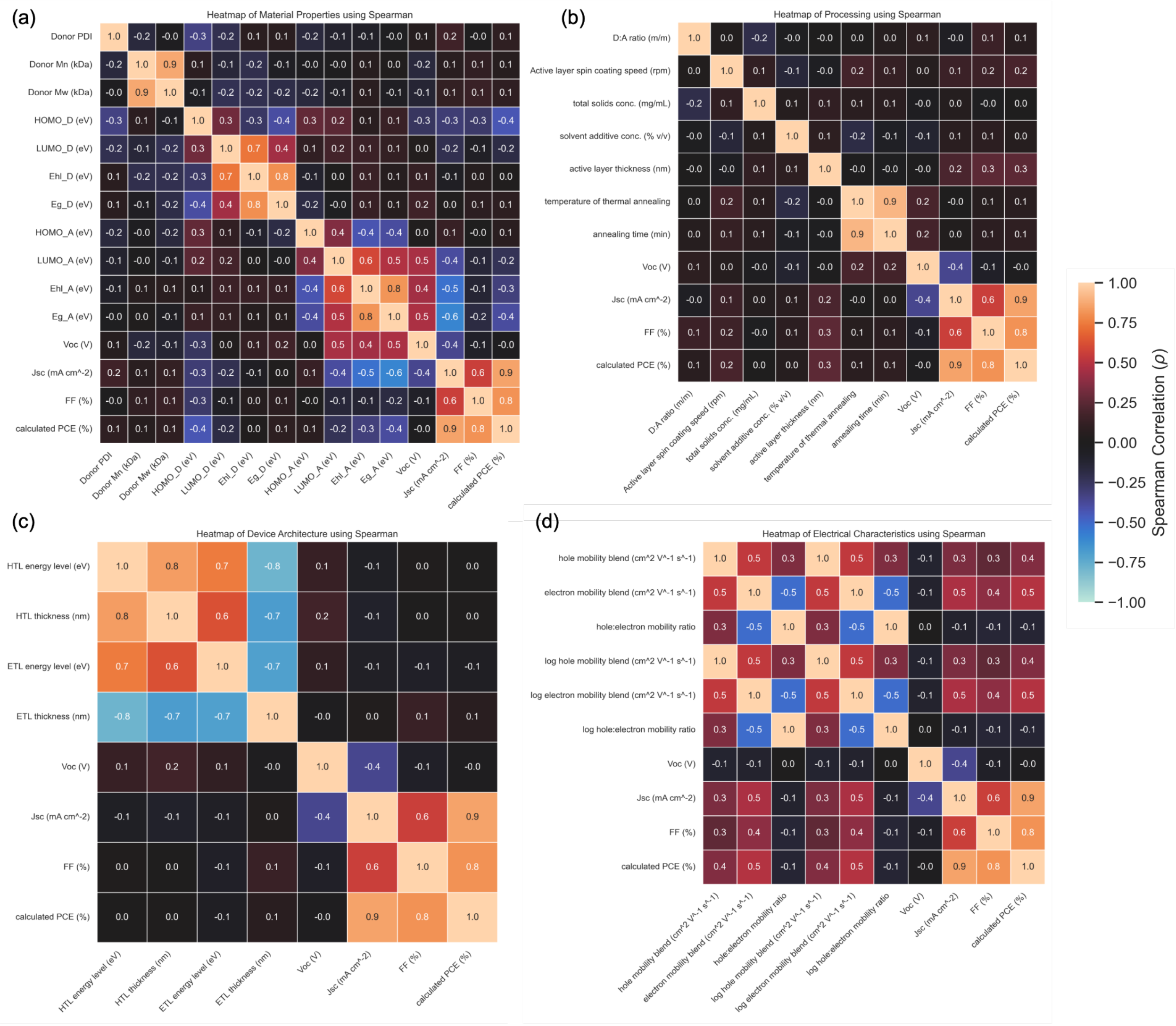

Figure S6. Spearman correlation coefficient heatmaps of the different groups of processing features with the output variables: (a) material properties, (b) processing, (c) device architecture, (d) charge carrier mobilities. Feature groups correspond to those defined in Figure S4. Correlation coefficient values in each cell are rounded to the nearest tenth for readability.

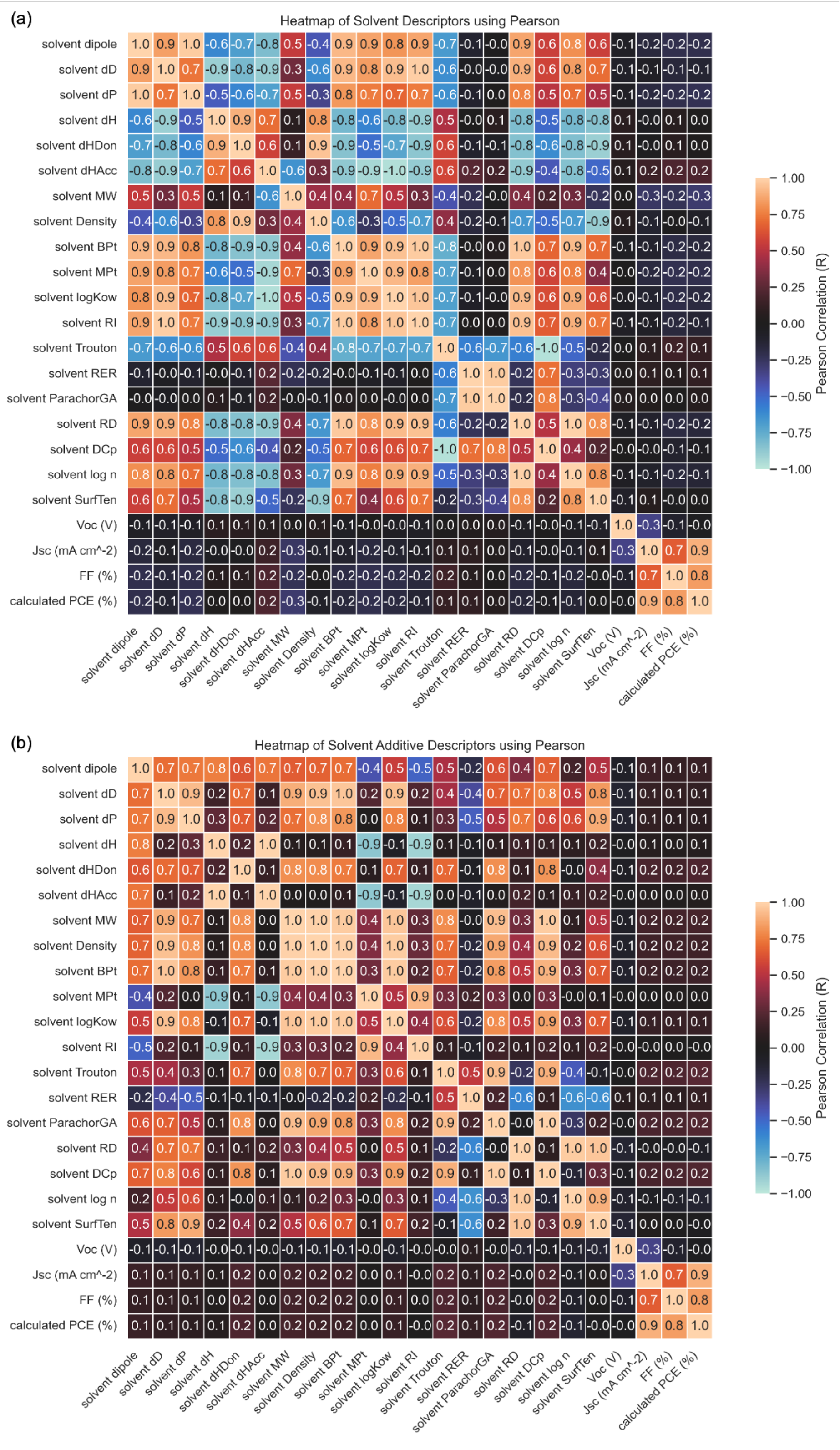

Figure S7. Pearson correlation coefficient heatmaps of (a) solvent and (b) solvent additive descriptors with the output variables. Correlation coefficient values in each cell are rounded to the nearest tenth for readability.

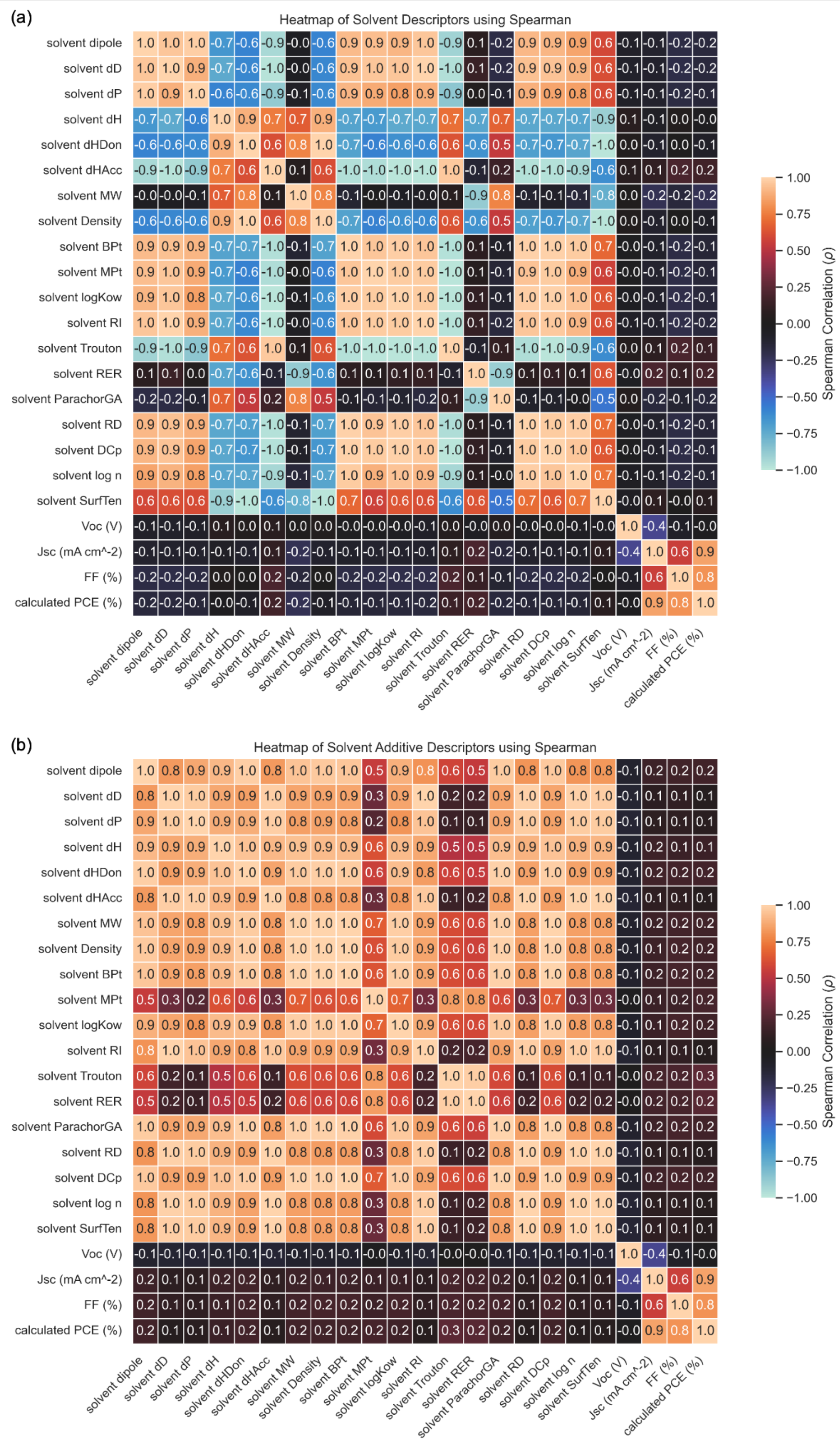

Figure S8. Spearman correlation coefficient heatmaps of (a) solvent and (b) solvent additive descriptors with the output variables. Correlation coefficient values in each cell are rounded to the nearest tenth for readability.

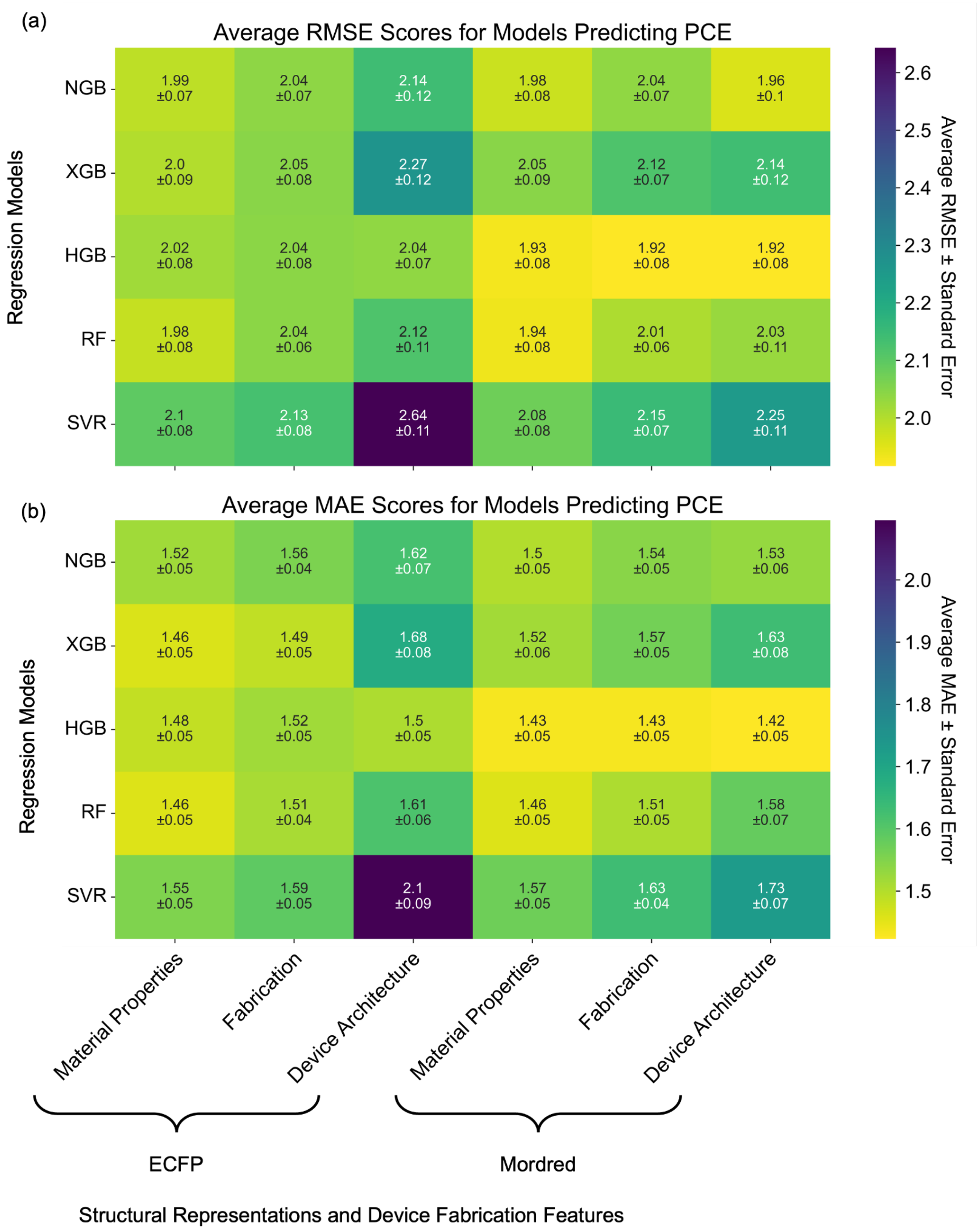

Figure S9. Heatmaps of model performance as measured by (a) RMSE and (b) MAE scores for predicting PCE from molecular structure and various subsets of OPV device fabrication parameters.

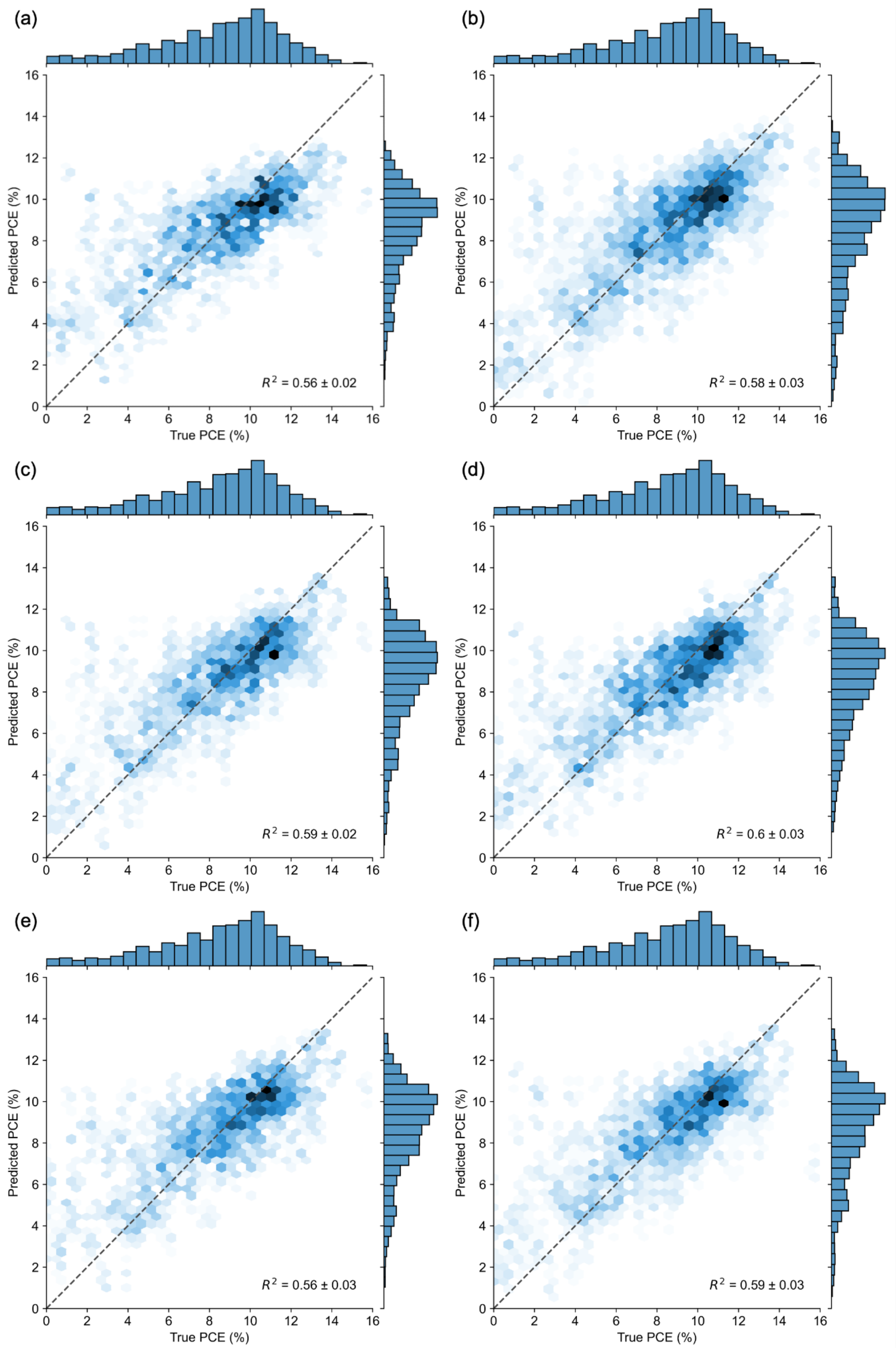

Figure S10. Scatter plots of predictions from (a) GP trained on ECFP only, (b) HGB trained on ECFP only, (c) RF trained on ECFP only, (d) RF trained on Mordred descriptors, (e) SVR trained on Mordred descriptors, (f) RF trained on ECFP and material properties.

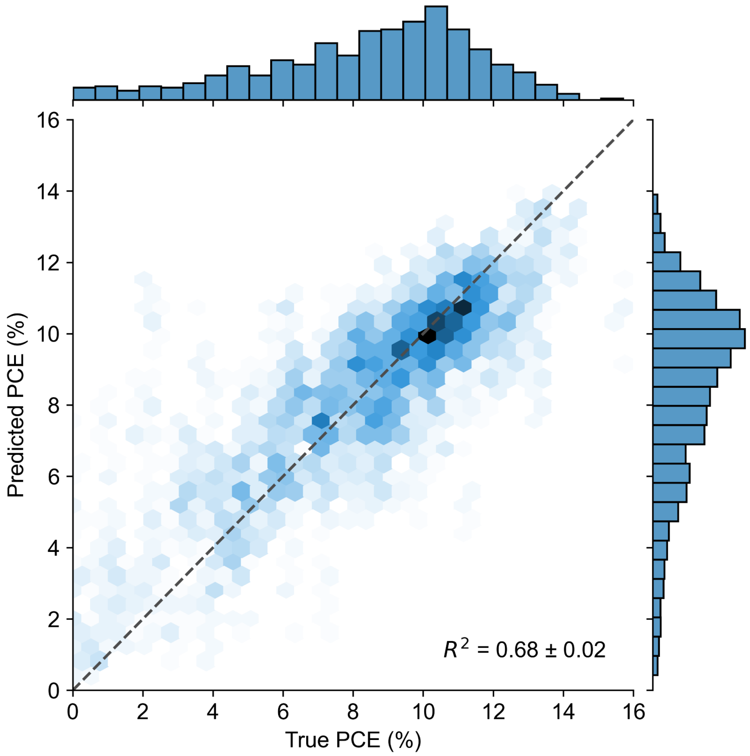

Figure S11. HGB trained on Mordred descriptors and log mobilities all subset.

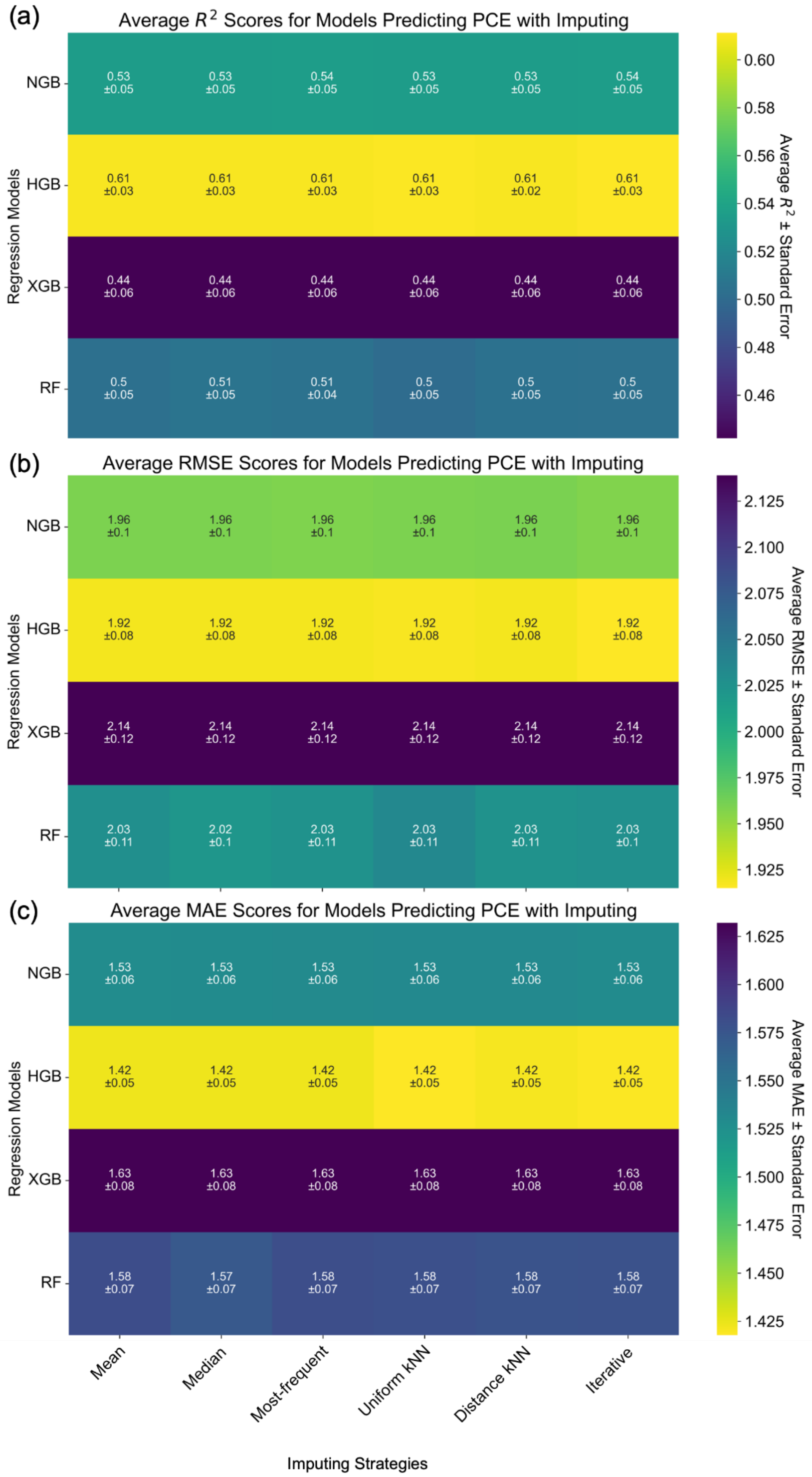

Figure S12. Heatmaps of model performance as measured by the (a) R^2, (b) RMSE, (c) MAE scores for predicting PCE from molecular structure and the device architecture subset of features when imputing missing data using different algorithms. Numbers in each cell correspond to the average score of the model over seven independent five-fold cross-validations ± the standard error of the mean.

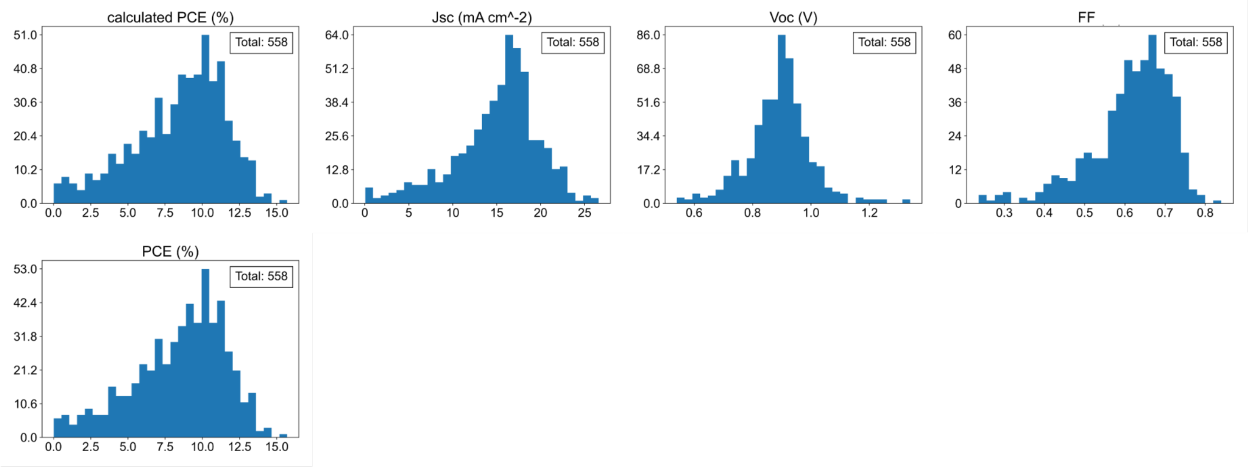

Figure S13. Distributions of target values in the dataset.

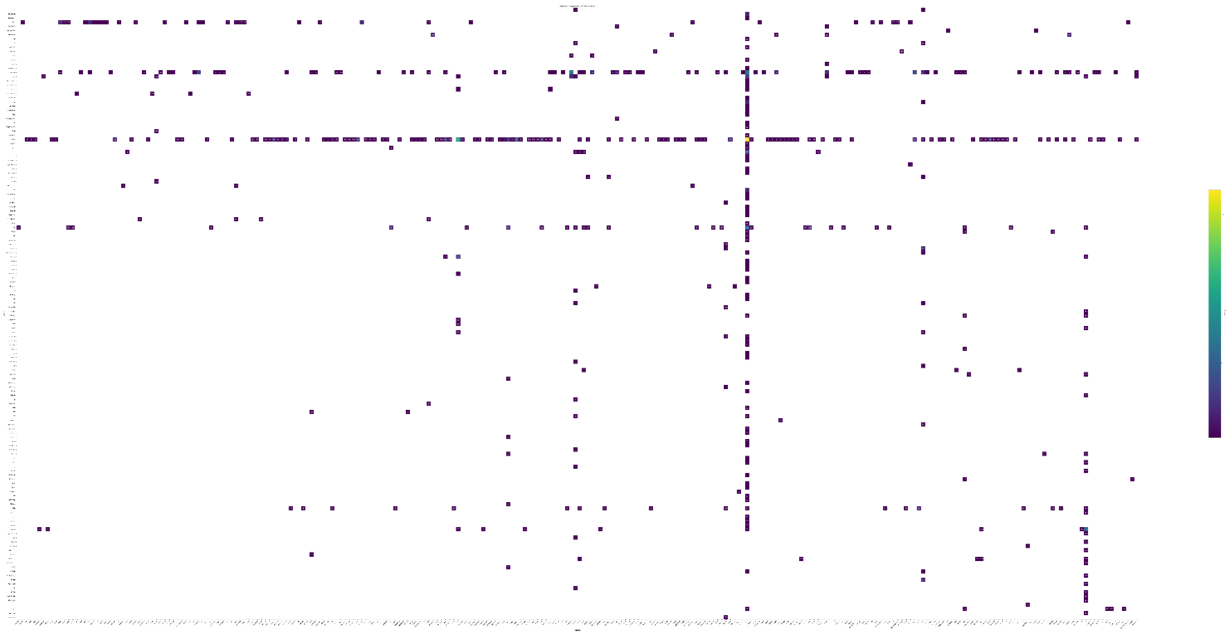

Figure S14. D:A heatmap. Most common donors: PTB7-Th, J71, P3HT, PBDB-T. Most common acceptors: ITIC, IT-4F, m-ITIC. X: more positive left, more negative right. Y: more positive top, more negative bottom.

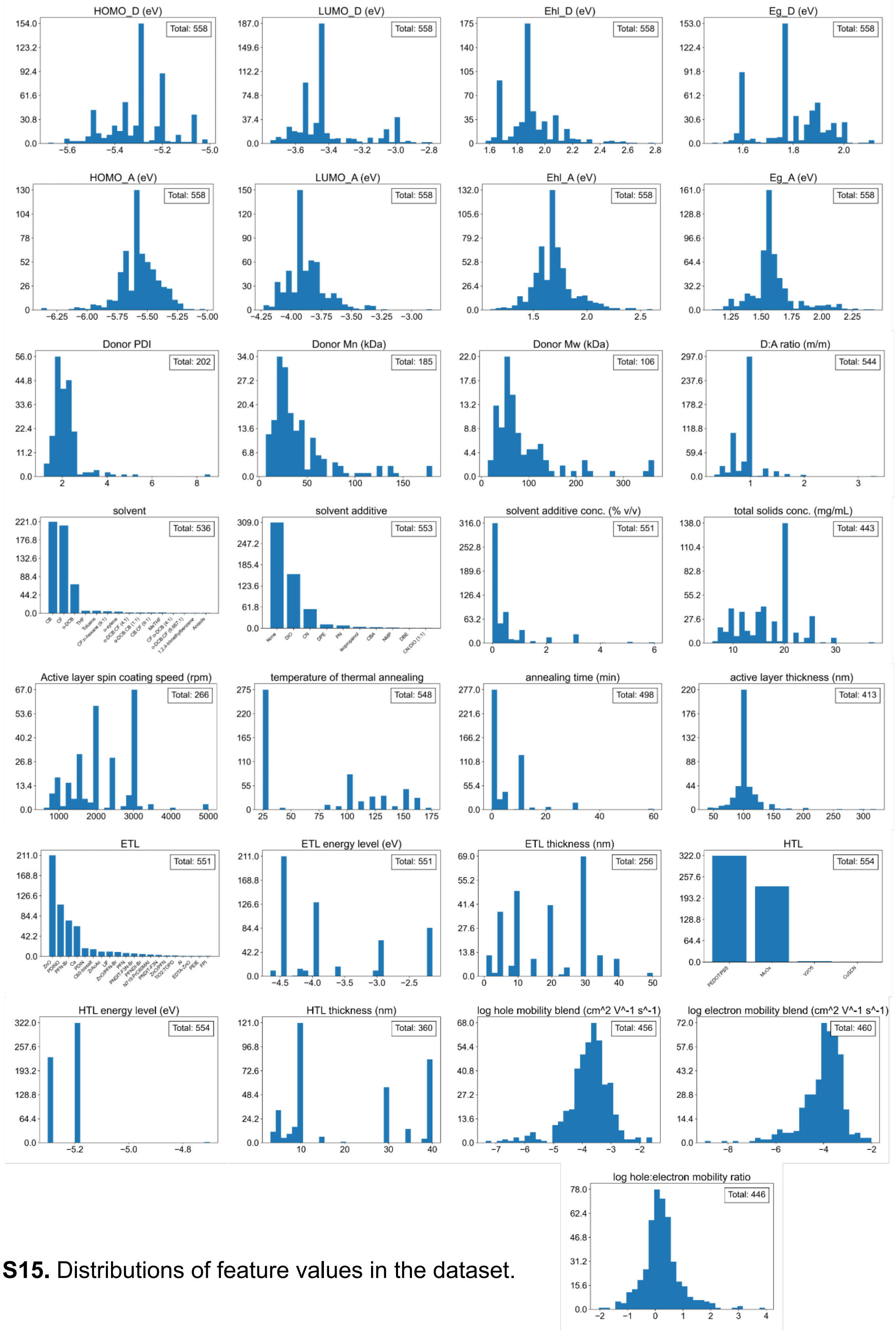

Figure S15. Distributions of feature values in the dataset.

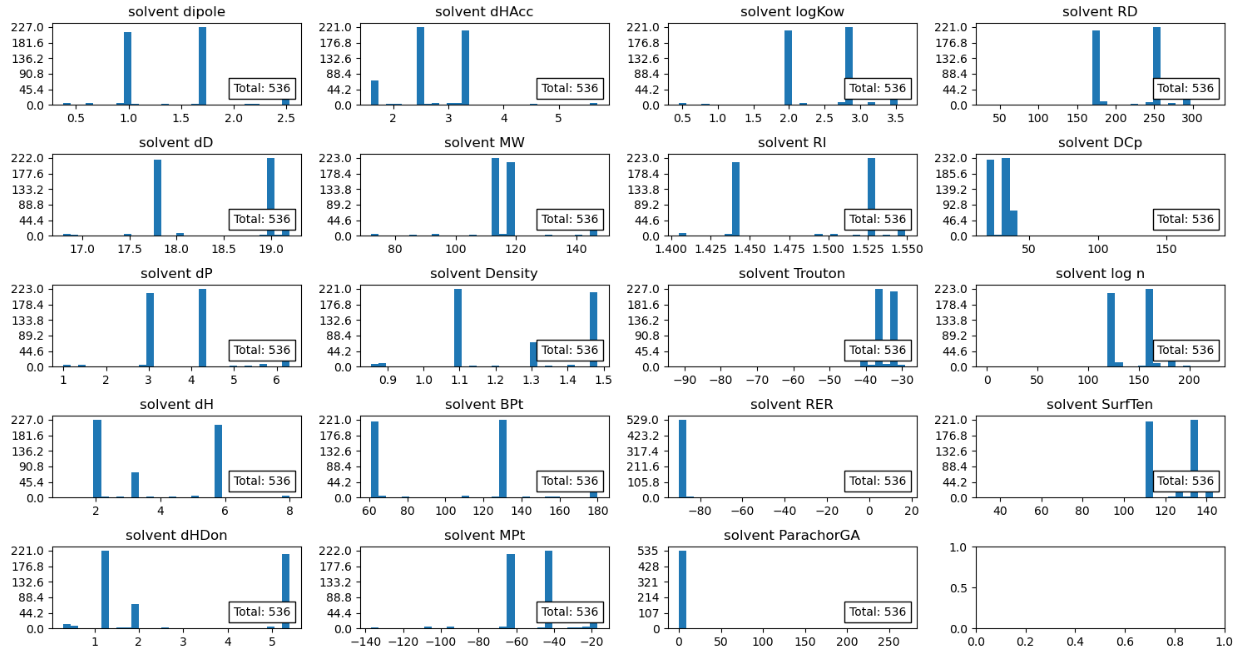
solvent dipole
solvent dHAcc
solvent logKow
solvent RD
solvent dD
solvent MW
solvent RI
solvent DCp
solvent dP
solvent Density
solvent Trouton
solvent log n
solvent dH
solvent BPt
solvent RER
solvent SurfTen
solvent dHDon
solvent MPt
solvent ParachorGA

Figure S16. Distributions of solvent descriptor values.

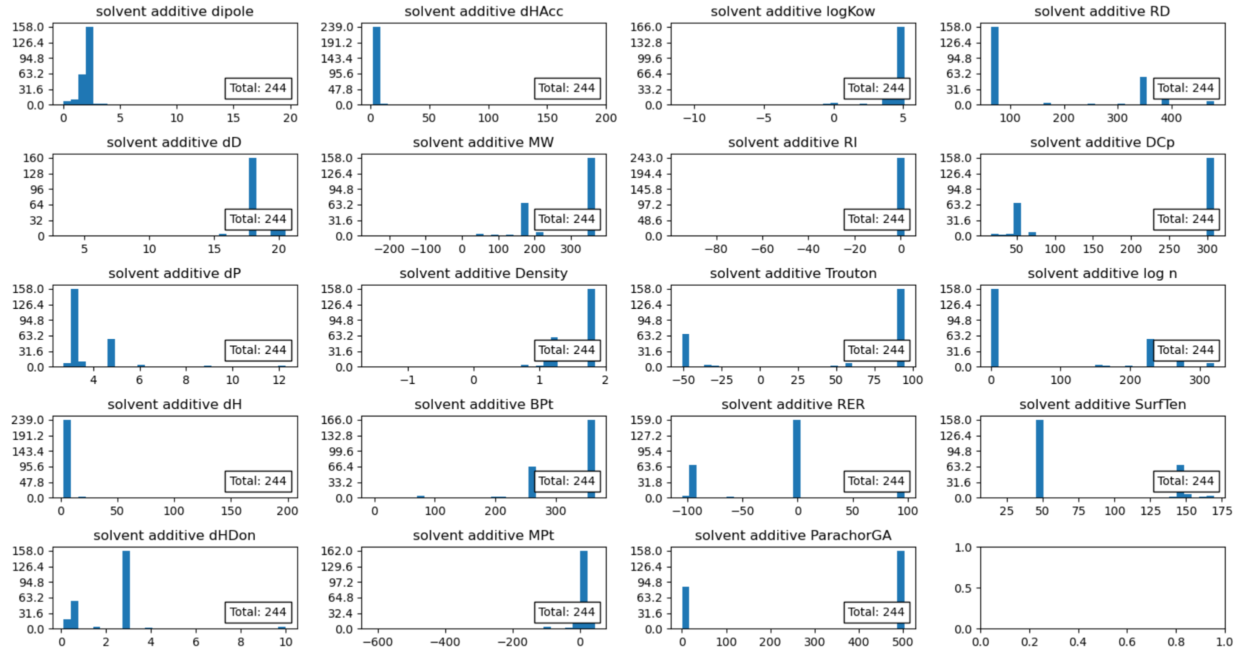
solvent additive dipole
solvent additive dHAcc
solvent additive logKow
solvent additive RD
solvent additive dD
solvent additive MW
solvent additive RI
solvent additive DCp
solvent additive dP
solvent additive Density
solvent additive Trouton
solvent additive log n
solvent additive dH
solvent additive BPt
solvent additive RER
solvent additive SurfTen
solvent additive dHDon
solvent additive MPt
solvent additive ParachorGA

Figure S17. Distributions of solvent additive descriptor values.

Table S5. Statistical analysis of feature distributions in the dataset.

Feature	Mean	Std Dev	Median	Mode	Skew	Min	Max	Missing (%)
Donor PDI	2.15	0.72	2	2	4.6	1.18	8.59	63.8
Donor Mn (kDa)	41.21	32.79	30	23.5	2.22	6.97	180.4	66.85
Donor Mw (kDa)	89.59	73.36	63.6	50.3	2.21	13.4	364.9	81
HOMO_D (eV)	-5.3	0.12	-5.3	-5.3	-0.01	-5.68	-5.01	0
LUMO_D (eV)	-3.4	0.2	-3.43	-3.43	1.11	-3.75	-2.78	0
Ehl_D (eV)	1.9	0.18	1.87	1.87	1.25	1.57	2.79	0
Eg_D (eV)	1.79	0.13	1.77	1.77	-0.17	1.48	2.12	0
HOMO_A (eV)	-5.56	0.16	-5.58	-5.58	-0.65	-6.38	-5.02	0
LUMO_A (eV)	-3.87	0.16	-3.91	-3.92	1.29	-4.23	-2.83	0
Ehl_A (eV)	1.69	0.19	1.66	1.66	0.89	1.1	2.61	0
Eg_A (eV)	1.6	0.18	1.59	1.59	1.21	1.1	2.43	0
D:A ratio (m/m)	0.92	0.26	1	1	2.04	0.33	3.33	2.51
Active layer spin coating speed (rpm)	2158.27	795.79	2000	3000	0.35	600	5000	52.33
total solids conc. (mg/mL)	16.8	5.13	18	20	0.16	6	37.5	20.61
solvent additive conc. (% v/v)	0.44	0.86	0	0	3.18	0	6	1.25
active layer thickness (nm)	103.61	25.17	100	100	3.77	40	321	25.99

Feature	Mean	Std Dev	Median	Mode	Skew	Min	Max	Missing (%)
Annealing temperature (ºC)	73.01	51.6	25	25	0.36	25	175	1.79
Annealing time (min)	4.66	7.51	0	0	2.92	0	60	10.75
HTL energy level (eV)	-5.24	0.06	-5.2	-5.2	2.5	-5.3	-4.7	0.72
HTL thickness (nm)	20.63	14.04	10	10	0.32	3	40	35.48
ETL energy level (eV)	-3.73	0.83	-3.9	-4.46	0.85	-4.65	-2.14	1.25
ETL thickness (nm)	18.92	11.6	20	30	0.2	0.7	50	54.12
hole mobility blend ($cm^2V^{-1}s^{-1}$)	0	0	0	0	9.86	0	0.03	18.28
electron mobility blend ($cm^2V^{-1}s^{-1}$)	0	0	0	0	7.59	0	0.01	17.56
V_{OC} (V)	0.89	0.11	0.9	0.9	0.03	0.54	1.34	0
J_{SC} ($mAcm^{-2}$)	15.07	4.74	15.96	16.8	-0.76	0.05	26.6	0
FF	0.62	0.1	0.63	0.65	-1.13	0.24	0.84	0
PCE (%)	8.44	3.09	9.05	6.4	-0.64	0.01	15.7	0
hole:electron mobility ratio	33.61	440.17	1.52	2	19.67	0.01	9090.91	20.07
log hole mobility blend ($cm^2V^{-1}s^{-1}$)	-3.79	0.76	-3.7	-3.82	-1.13	-7.36	-1.57	18.28
log electron	-4.02	0.85	-3.87	-3.92	-1.54	-8.96	-2.01	17.56

Feature	Mean	Std Dev	Median	Mode	Skew	Min	Max	Missing (%)
mobility blend (cm^{2}V^{-1}s^{-1})								
log hole:electron mobility ratio	0.23	0.69	0.18	0.3	0.9	-2.04	3.96	20.07
calculated PCE (%)	8.43	3.09	9.02	4.5	-0.64	0.01	15.71	0

Table S6. Performance of HGB models trained on ECFP molecular representations with different radii and numbers of bits.

Representation	Radius	Bits	R^2	RMSE	MAE
ECFP6	3	512	0.56±0.02	2.05±0.07	1.53±0.05
ECFP8	4	1024	0.57±0.03	2.02±0.08	1.50±0.05
ECFP10	5	2048	0.58±0.03	2.00±0.08	1.48±0.05
ECFP12	6	4096	0.58±0.03	1.99±0.08	1.49±0.05

References

1. Greenstein, B. L. & Hutchison, G. R. Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms. *J. Phys. Chem. C* **127**, 6179–6191 (2023).
2. Miyake, Y. & Saeki, A. Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J. Phys. Chem. Lett.* **12**, 12391–12401 (2021).
3. Li, C.-Z. *et al.* Effective interfacial layer to enhance efficiency of polymer solar cells via solution-processed fullerene-surfactants. *J. Mater. Chem.* **22**, 8574–8578 (2012).
4. Allen, T. G. *et al.* Calcium contacts to n-type crystalline silicon solar cells. *Prog. Photovolt. Res. Appl.* **25**, 636–644 (2017).
5. Li, X., Liu, X., Zhang, W., Wang, H.-Q. & Fang, J. Fullerene-Free Organic Solar Cells with Efficiency Over 12% Based on EDTA–ZnO Hybrid Cathode Interlayer. *Chem. Mater.* **29**, 4176–4180 (2017).
6. Karki, A. *et al.* The role of bulk and interfacial morphology in charge generation, recombination, and extraction in non-fullerene acceptor organic solar cells. *Energy Environ. Sci.* **13**, 3679–3692 (2020).
7. Li, W. *et al.* A New Function of N719: N719 Based Solution-Processible Binary Cathode Buffer Layer Enables High-Efficiency Single-Junction Polymer Solar Cells. *Sol. RRL* **1**, 1700014 (2017).
8. Zhang, Y., Chen, L., Hu, X., Zhang, L. & Chen, Y. Low Work-function Poly(3,4-ethylenedioxylenethiophene): Poly(styrene sulfonate) as Electron-transport Layer for High-efficient and Stable Polymer Solar Cells. *Sci. Rep.* **5**, 12839 (2015).
9. Huang, F. *et al.* High-Efficiency, Environment-Friendly Electroluminescent Polymers with

Stable High Work Function Metal as a Cathode: Green- and Yellow-Emitting Conjugated Polyfluorene Polyelectrolytes and Their Neutral Precursors. *J. Am. Chem. Soc.* **126**, 9845–9853 (2004).

10. Yang, T. *et al.* Inverted polymer solar cells with 8.4% efficiency by conjugated polyelectrolyte. *Energy Environ. Sci.* **5**, 8208–8214 (2012).
11. Wu, Z. *et al.* n-Type Water/Alcohol-Soluble Naphthalene Diimide-Based Conjugated Polymers for High-Performance Polymer Solar Cells. *J. Am. Chem. Soc.* **138**, 2004–2013 (2016).
12. Yan, Y. *et al.* Light-Soaking-Free Inverted Polymer Solar Cells with an Efficiency of 10.5% by Compositional and Surface Modifications to a Low-Temperature-Processed TiO2 Electron-Transport Layer. *Adv. Mater.* **29**, 1604044 (2017).
13. Schopp, N. *et al.* Understanding Interfacial Recombination Processes in Narrow-Band-Gap Organic Solar Cells. *ACS Energy Lett.* **7**, 1626–1634 (2022).
14. Tan, Z. *et al.* High performance polymer solar cells with as-prepared zirconium acetylacetonate film as cathode buffer layer. *Sci. Rep.* **4**, 4691 (2014).
15. Montgomery, A. *et al.* Solution-processed copper (I) thiocyanate (CuSCN) for highly efficient CdSe/CdTe thin-film solar cells. *Prog. Photovolt. Res. Appl.* **27**, 665–672 (2019).
16. Li, X., Xie, F., Zhang, S., Hou, J. & Choy, W. C. MoOx and V2Ox as hole and electron transport layers through functionalized intercalation in normal and inverted organic optoelectronic devices. *Light Sci. Appl.* **4**, e273–e273 (2015).
17. Chen, S. *et al.* Efficient Nonfullerene Organic Solar Cells with Small Driving Forces for Both Hole and Electron Transfer. *Adv. Mater.* **30**, 1804215 (2018).
18. Abbott, S. HSPiP: Hansen Solubility Parameters in Practice.