Supporting Information

Identifying structure-absorption relationships and predicting absorption strength of non-fullerene acceptors for organic photovoltaics

Jun Yan,[a,#] Xabier Rodríguez-Martínez,[b,c,#] Drew Pearce,[a] Hana Douglas,[a] Danai Bili,[a] Mohammed Azzouzi,[a] Flurin Eisner,[a] Alise Virbule,[a] Elham Rezasoltani,[a] Valentina Belova,[c] Bernhard Dörling,[c] Sheridan Few,[a,f] Anna A. Szumska,[a] Xueyan Hou,[a] Guichuan Zhang,[d] Hin-Lap Yip,[d,e] Mariano Campoy-Quiles *[,c] and Jenny Nelson*[,a]*

[#] J.Y. and X.R.-M. contributed equally to this work.

[a] Department of Physics, Imperial College London, SW7 2AZ, London, United Kingdom
Email: jenny.nelson@imperial.ac.uk

[b] Electronic and Photonic Materials (EFM), Department of Physics, Chemistry and Biology (IFM), Linköping University, Linköping, SE 581 83 Sweden
Email: xabier.rodriguez.martinez@liu.se

[c] Instituto de Ciencia de Materiales de Barcelona, ICMAB-CSIC, Campus UAB, Bellaterra 08193, Spain
Email: mcampoy@icmab.es

[d] Institute of Polymer Optoelectronic Materials and Devices, State Key Laboratory of Luminescent Materials and Devices, South China University of Technology, Guangzhou 510640, P. R. China

[e] Department of Materials Science and Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

[f] Sustainability Research Institute, School of Earth and Environment, University of Leeds, Leeds, LS2 9JT

Table of Contents

Supplementary Note 1

Supplementary Note 1. Chemical names and nomenclature of the materials highlighted in this work.

PC61BM: [6,6]-Phenyl-C_{61}-butyric acid methyl ester

PC71BM: [6,6]-Phenyl-C_{71}-butyric acid methyl ester

ICBA: 1′,1″,4′,4″-Tetrahydro-di[1,4]methanonaphthaleno[1,2:2′,3′,56,60:2″,3″][5,6]fullerene-C_{60}

Y5: (2,2'-((2Z,2'Z)-((12,13-bis(2-ethylhexyl)-3,9-diundecyl-12,13-dihydro[1,2,5]thiadiazolo[3,4e]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g] thieno[2',3':4,5]thieno[3,2-b]-indole-2,10-diyl)bis(methanylylidene))bis(3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile)

Y6: 2,2'-((2Z,2'Z)-((12,13-bis(2-ethylhexyl)-3,9-diundecyl-12,13-dihydro-[1,2,5]thiadiazolo[3,4-e]thieno[2",3'':4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methanylylidene))bis(5,6-difluoro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

Y7: 2,2'-((2Z,2'Z)-((12,13-bis(2-ethylhexyl)-3,9-diundecyl-12,13-dihydro-[1,2,5]thiadiazolo[3,4-e]thieno[2",3'':4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methanylylidene))bis(5,6-dichloro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

Y11: 2,2'-((2Z,2'Z)-((6,12,13-tris(2-ethylhexyl)-3,9-diundecyl-12,13-dihydro-6H-thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b][1,2,3]triazolo[4,5-e]indole-2,10-diyl)bis(methanylylidene))bis(5,6-difluoro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

Y12: 2,2'-((2Z,2'Z)-((12,13-bis(2-butyloctyl)-3,9-diundecyl-12,13-dihydro-[1,2,5]thiadiazolo[3,4-e]thieno[2",3'':4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methanylylidene))bis(5,6-difluoro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

O-IDTBR: (5Z,5'Z)-5,5'-(((4,4,9,9-tetrakis(n-octyl)-4,9-dihydro-s-indaceno[1,2-b:5,6-b']dithiophene-2,7-diyl)bis(benzo[c][1,2,5]thiadiazole-7,4-diyl))bis(methaneylylidene))bis(3-ethyl-2-thioxothiazolidin-4-one)

O-IDFBR: (5Z,5'Z)-5,5'-(((6,6,12,12-tetraoctyl-6,12-dihydroindeno[1,2-b]fluorene-2,8-diyl)bis(benzo[c][1,2,5]thiadiazole-7,4-diyl))bis(methaneylylidene))bis(3-ethyl-2-thioxothiazolidin-4-one)

EH-IDTBR: (5Z,5'Z)-5,5'-(((4,4,9,9-tetrakis(2-ethylhexyl)-4,9-dihydro-s-indaceno[1,2-b:5,6-b']dithiophene-2,7-diyl)bis(benzo[c][1,2,5]thiadiazole-7,4-diyl))bis(methaneylylidene))bis(3-ethyl-2-thioxothiazolidin-4-one)

IDIC: 2,2'-((2Z,2'Z)-((4,4,9,9-tetrahexyl-4,9-dihydro-s-indaceno[1,2-b:5,6-b']dithiophene-2,7-diyl)bis(methanylylidene))bis(3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

SN6IC-4F: 2,2'-((2Z,2'Z)-((thieno[3,2-b]thieno[2''',3''':4'',5'']pyrrolo[2'',3'':4',5']thieno[2',3':4,5]thieno[2,3- d]pyrrole,4,9-dihydro-4,9-di-1-octylnonyl-2,7-diyl)bis(methanylylidene))bis((5,6-difluoro-3-oxo-2,3 -dihydro-1H-indene-2,1-diylidene))dimalononitrile

ITIC: 2,2'-[[6,6,12,12-tetrakis(4-hexylphenyl)-6,12-dihydrodithieno[2,3-d:2',3'-d']-s-indaceno[1,2-b:5,6-b']dithiophene-2,8-diyl]bis[methylidyne(3-oxo-1H-indene-2,1(3H)-diylidene)]]bis[propanedinitrile]

ITIC-C_2C_6: 2,2'-[[6,6,12,12-tetrakis(2-ethylhexyl)-6,12-dihydrodithieno[2,3-d:2',3'-d']-s-indaceno[1,2-b:5,6-b']dithiophene-2,8-diyl]bis[methylidyne(3-oxo-1H-indene-2,1(3H)-diylidene)]]bis[propanedinitrile]

ITIC-C_8: 2,2'-[[6,6,12,12-tetrakis(n-octyl)-6,12-dihydrodithieno[2,3-d:2',3'-d']-s-indaceno[1,2-b:5,6-b']dithiophene-2,8-diyl]bis[methylidyne(3-oxo-1H-indene-2,1(3H)-diylidene)]]bis[propanedinitrile]

IT-4F: 9-Bis(2-methylene-((3-(1,1-dicyanomethylene)-6,7-difluoro)-indanone))-5,5,11,11-tetrakis(4-hexylphenyl)-dithieno[2,3-d:2',3'-d']-s-indaceno[1,2-b:5,6-b']dithiophene

CBM: 2,2'-(7,7'-(9-(heptadecan-9-yl)-9H-carbazole-2,7-diyl)bis(benzo[c][1,2,5]thiadiazole-7,4-diyl))bis(methan-1-yl-1-ylidene)dimalononitrile

FBR: 5,5′-[(9,9-Dioctyl-9H-fluorene-2,7-diyl)bis(2,1,3-benzothiadiazole-7,4-diylmethylidyne)]bis[3-ethyl-2-thioxo-4-thiazolidinone]

BTPM: 2,2'-((2Z,2'Z)-((12,13-diisobutyl-3,9-dimethyl-5,7,12,13-tetrahydro-[1,2,5]thiadiazolo[3,4-e]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methaneylylidene))bis(6-methyl-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

BTTPC: 2,2'-((6Z,6'Z)-((12,13-diisobutyl-3,9-dimethyl-5,7,12,13-tetrahydro-[1,2,5]thiadiazolo[3,4-e]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methaneylylidene))bis(5-oxo-5,6-dihydro-7H-indeno[5,6-b]thiophene-6,7-diylidene))dimalononitrile

BTDTP-4F: 2,2'-((2Z,2'Z)-((3,12-dimethyl-13,14-dihydro-12H-[1,2,5]thiadiazolo[3,4-e]thieno[2",3":4',5']pyrrolo[2',3':4,5]thieno[3,2-b]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]indole-2,10-diyl)bis(methaneylylidene))bis(5,6-difluoro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

BDTP-4F: 2,2'-(((1Z,1'Z)-(1,11-dimethyl-4,6,6c,10,11,11b,12,13-octahydro-2H-[1,2,5]thiadiazolo[3,4-e]thieno[2",3":4',5']pyrrolo[2',3':4,5]thieno[3,2-b]thieno[2',3':4,5]pyrrolo[3,2-g]indole-2,9(1H)-diylidene)bis(methaneylylidene))bis(5,6-difluoro-3-oxo-2,3,3a,6,7,7a-hexahydro-1H-indene-2-yl-1-ylidene))dimalononitrile

BTTPTP-2OYPD: 2,2'-((2Z,2'Z)-((13,14-diisobutyl-5,7,13,14-tetrahydro-[1,2,5]thiadiazolo[3,4-e]thieno[2"',3"':4",5"]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2",3":4',5']thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methaneylylidene))bis(3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

BTPTTT-2OYPD: 2,2'-((2Z,2'Z)-((13,14-diisobutyl-5,7,13,14-tetrahydro-[1,2,5]thiadiazolo[3,4-e]thieno[2"',3"':4",5"]thieno[2",3":4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2",3":4',5']thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methaneylylidene))bis(3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

IEICO: 2,2′-((2Z,2′Z)-((5,5′-(4,4,9,9-tetrakis(4-hexylphenyl)-4,9-dihydros-indaceno[1,2-b:5,6-b′]dithiophene-2,7-diyl)bis(4-((2-ethylhexyl)oxy)thiophene-5,2-diyl))bis(methanylylidene))bis(3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

IEICO-4F: 2,2'-((2Z,2'Z)-(((4,4,9,9-tetrakis(4-hexylphenyl)-4,9-dihydro-sindaceno[1,2-b:5,6-b']dithiophene-2,7-diyl)bis(4-((2-ethylhexyl)oxy)thiophene-5,2-diyl))bis(methanylylidene))bis(5,6-difluoro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

BTTPTP-4Cl: 2,2'-((2Z,2'Z)-((13,14-diisobutyl-5,7,13,14-tetrahydro-[1,2,5]thiadiazolo[3,4-e]thieno[2''',3''':4'',5'']thieno[2'',3'':4',5']thieno[2',3':4,5]pyrrolo[3,2-g]thieno[2'',3'':4',5']thieno[2',3':4,5]thieno[3,2-b]indole-2,10-diyl)bis(methaneylylidene))bis(5,6-dichloro-3-oxo-2,3-dihydro-1H-indene-2,1-diylidene))dimalononitrile

INPIC-4F: [(Z)-2-({24-[(Z)-(1-Dicyanomethylidene-5,6-difluoro-3-oxo-2-indanylidene)methyl]-15,15,30,30-tetrakis(p-hexylphenyl)-12,27-dioctyl-5,8,20,23-tetrathia-12,27-diazanonacyclo[16.12.0.$0^{3,16}$.$0^{4,14}$.$0^{6,13}$.$0^{7,11}$.$0^{19,29}$.$0^{21,28}$.$0^{22,26}$]triaconta-1(18),2,4(14),6(13),7(11),9,16,19(29),21(28),22(26),24-undecaen-9-yl}methylidene)-5,6-difluoro-3-oxo-1-indanylidene]propanedinitrile

o-IO1: 2-((Z)-2-((5-(7-(5-(((Z)-1-(dicyanomethylene)-5,6-difluoro-3-oxo-1,3-dihydro-2H-inden-2-ylidene)methyl)-3-((2-ethylhexyl)oxy)thiophen-2-yl)-4,4,9,9-tetraoctyl-4,9-dihydro-s-indaceno[1,2-b:5,6-b']dithiophen-2-yl)-4-(2-ethylhexyl)thiophen-2-yl)methylene)-5,6-difluoro-3-oxo-2,3-dihydro-1H-inden-1-ylidene)malononitrile

TfIF-4FIC: [(Z)-2-({26-[(Z)-(1-Dicyanomethylidene-5,6-difluoro-3-oxo-2-indanylidene)methyl]-7,7,16,16,23,23,32,32-octaoctyl-11,27-dithianonacyclo[17.13.0.0^{3,17}.$0^{4,15}$.$0^{6,13}$.$0^{8,12}$.$0^{20,31}$.$0^{22,29}$.$0^{24,28}$]dotriaconta-1(19),2,4(15),5,8(12),9,13,17,20(31),21,24(28),25,29-tridecaen-10-yl}methylidene)-5,6-difluoro-3-oxo-1-indanylidene]propanedinitrile

NIBT: (7Z,7'Z)-7,7'-(((4,4,9,9-tetrakis(4-octylphenyl)-4,9-dihydro-s-indaceno[1,2-b:5,6-b']dithiophene-2,7-diyl)bis(benzo[c][1,2,5]thiadiazole-7,4-diyl))bis(methaneylylidene))bis(2-(2-ethylhexyl)-1H-indeno[6,7,1-def]isoquinoline-1,3,6(2H,7H)-trione)

Figure S1

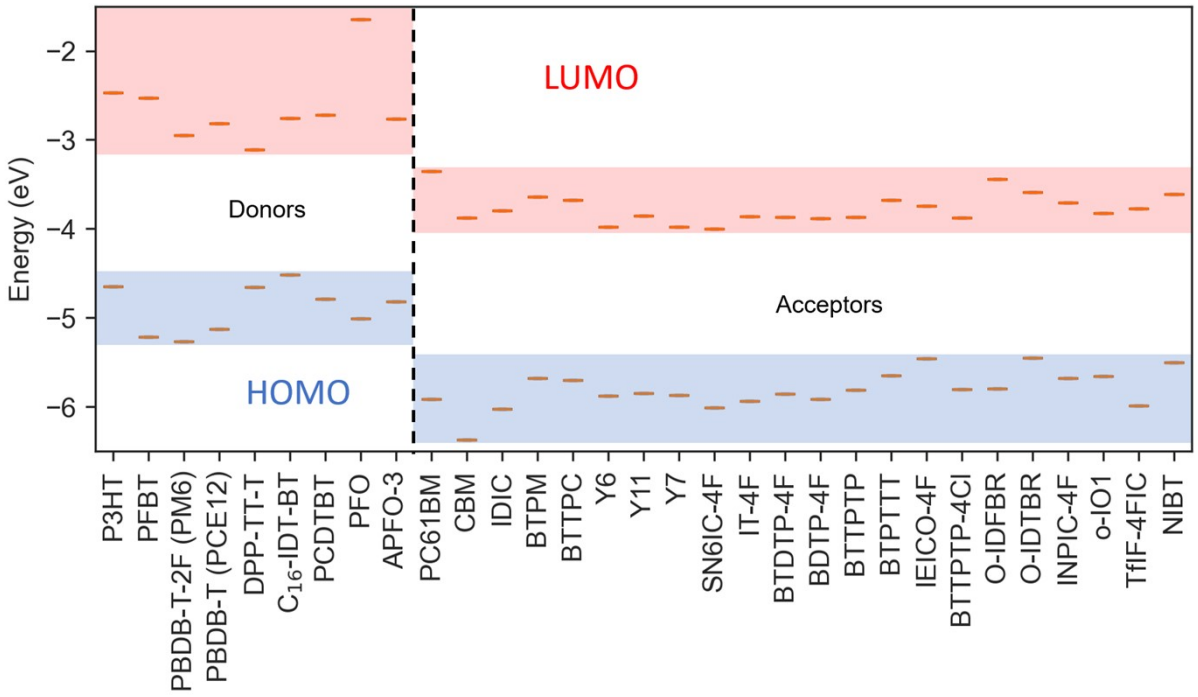


Figure S1. Highest occupied molecular orbital (HOMO) and lowest unocuppied molecular orbital (LUMO) energy levels of representative donor and acceptor molecules as retrieved from TDDFT calculations. Molecules considered as NFAs in this work show proper HOMO and LUMO energy level alignment to act as electron acceptor when in a bulk heterojunction blend with commonly used donors, such as P3HT, PCDTBT, PM6 or PBDB-T.

Figure S2

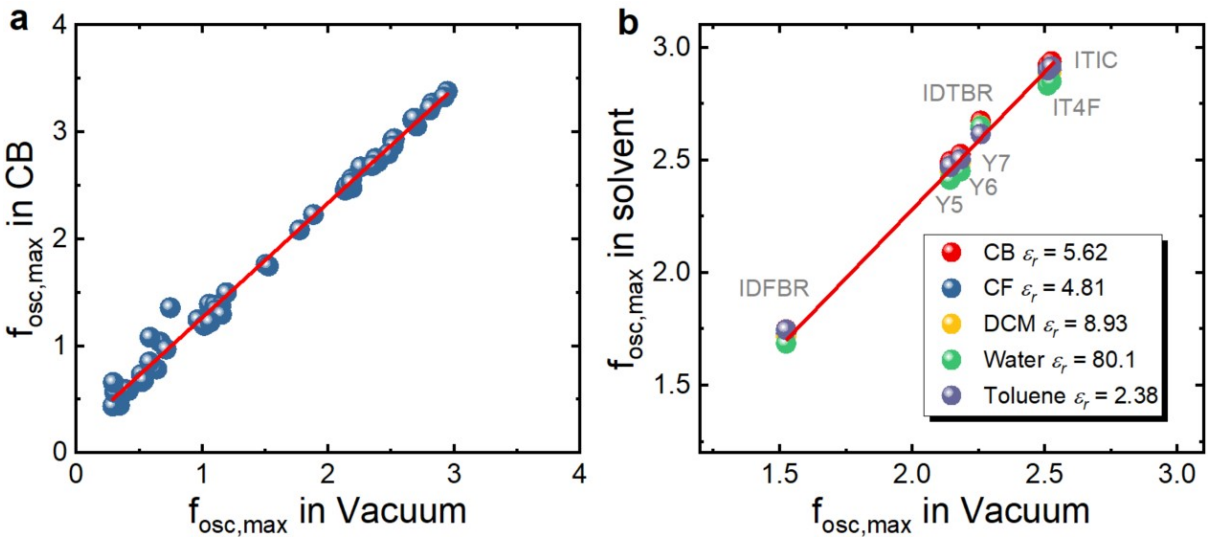


Figure S2. Solvent effect on the maximum oscillator strength ($f_{osc,max}$) in TDDFT calculations. (a) $f_{osc,max}$ in Chlorobenzene (CB) versus in vacuum for 56 organic molecules including common NFAs. (**b)** $f_{osc,max}$ in various organic solvents versus in vacuum for 7 different organic molecules, O-IDFBR, O-IDTBR, IT-4F, ITIC, Y5, Y6, and Y7. Noting here that $\varepsilon_{d,max} \propto f_{osc,max}$. TDDFT was performed under B3LYP/6-311+G(d,p) using Polarizable-continuum-solvent-model (PCM). We can see that the choice of solvent does not affect $f_{osc,max}$ much, and that a good linear correlation between solvent and vacuum $f_{osc,max}$ is obtained. This tells us that the same correlation between TDDFT and experiments will be maintained based on either vacuum environment or polarized medium, which allows us focus on TDDFT results from vacuum calculations only. This is a great benefit since most of the TDDFT calculations by the present and past group members were done in vacuum, allowing us to have a larger database.

Figure S3

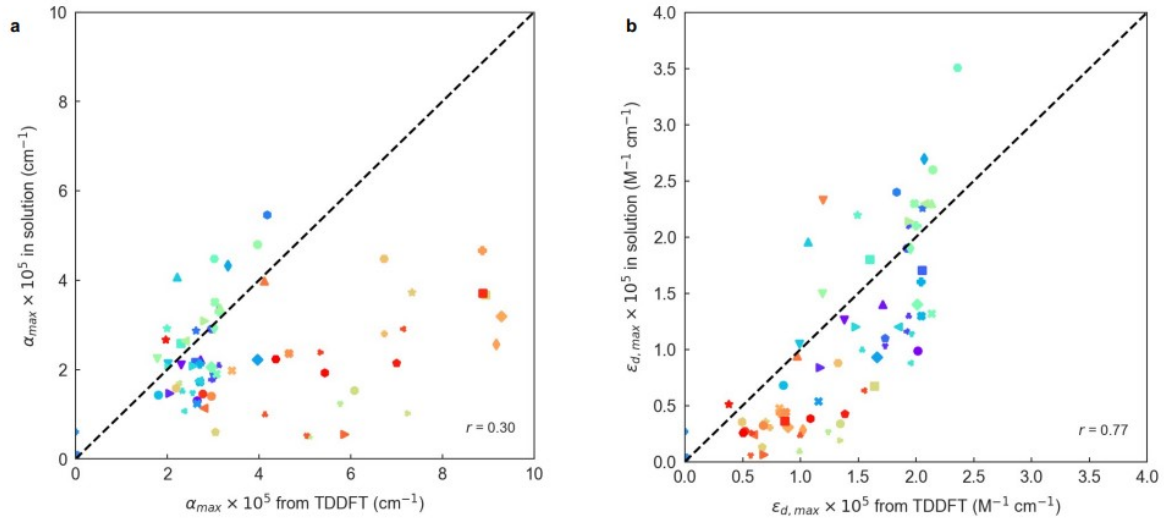


Figure S3. (a) Correlation between the maximum of the absorption coefficient ($\varepsilon_{d,max}$) obtained in solution state with the TDDFT calculated values. (b) Same dataset yet plotted in terms of maximum molar extinction coefficient $\varepsilon_{d,max}$). A significant fraction of this dataset was collected from literature.[1,2,11,12,3–10] When required, a refractive index of 1.5 and a solid density of 1000 g L^{-1} were considered for all materials.

Figure S4

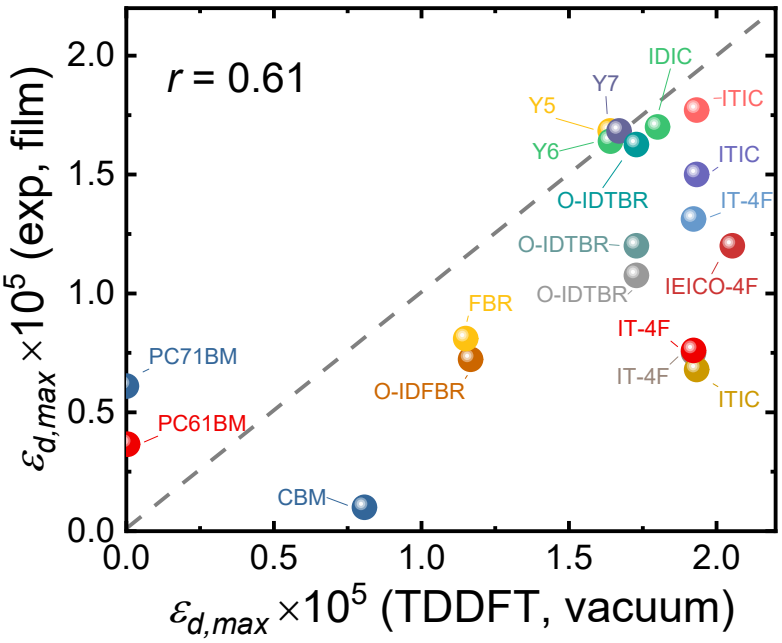


Figure S4. Comparison of maximum molar extinction coefficient ($\varepsilon_{d,max}$) between film experiments and TDDFT calculations in vacuum (20 data points). The experimental data of $\varepsilon_{d,max}$ in film are either measured or collected from literature, noting that different values may be present for the same material as they were collected from different papers. Unit of $\varepsilon_{d,max}$ is M^{-1} cm^{-1}. Grey dashed lines indicate the perfect match between x- and y-axis. The detailed data for generating this figure is presented in the Supplementary Database.

Figure S5

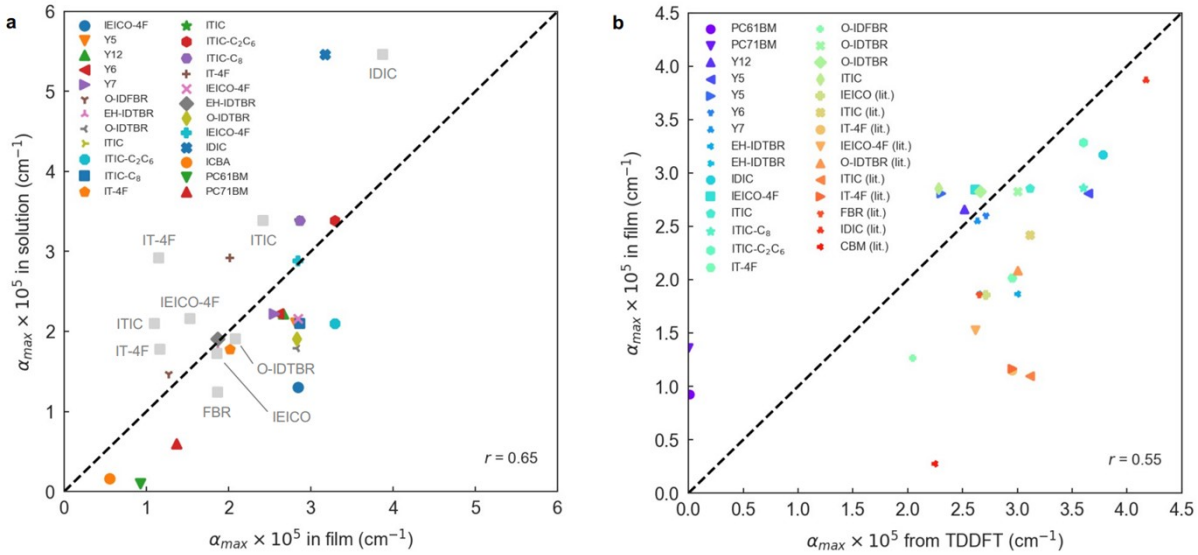


Figure S5. (a) Maximum absorption coefficient ($^{\alpha_{max}}$) in solution and film. Grey squares correspond to data obtained from literature.[1,11] (b) Maximum absorption coefficient in film and as obtained in their corresponding TDDFT calculations. A few data points (labelled as lit.) correspond to values extracted from literature.[1,11]

Figure S6

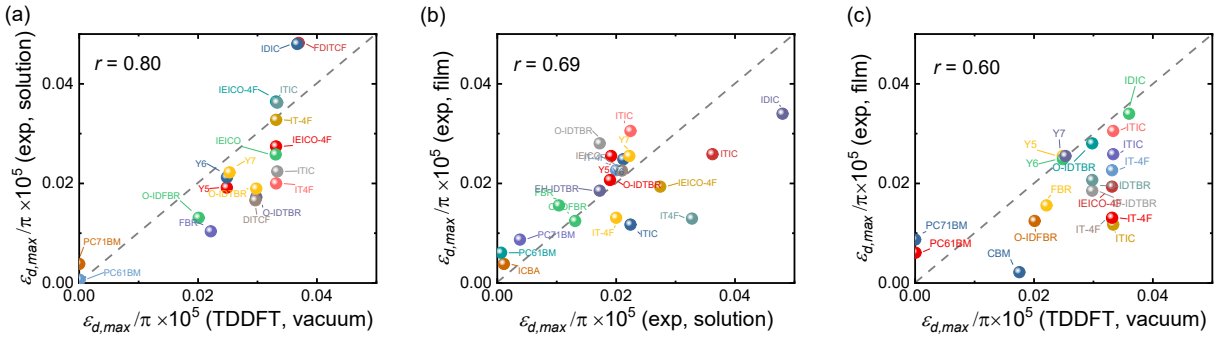


Figure S6. Effect of the number of π-electrons on the comparison between experimental (solution and film) maximum molar extinction coefficients and TDDFT results for the NFAs studied. (a) Experimental $\varepsilon_{d,max}/\pi$ in solution versus calculated $\varepsilon_{d,max}/\pi$ using TDDFT, (b) experimental $\varepsilon_{d,max}/\pi$ in film (solid state) versus that in solution; and (c) experimental $\varepsilon_{d,max}/\pi$ in solid state film versus calculated $\varepsilon_{d,max}/\pi$ using TDDFT. Unit of $\varepsilon_{d,max}/\pi$ is M^{-1} cm^{-1}.

Figure S7

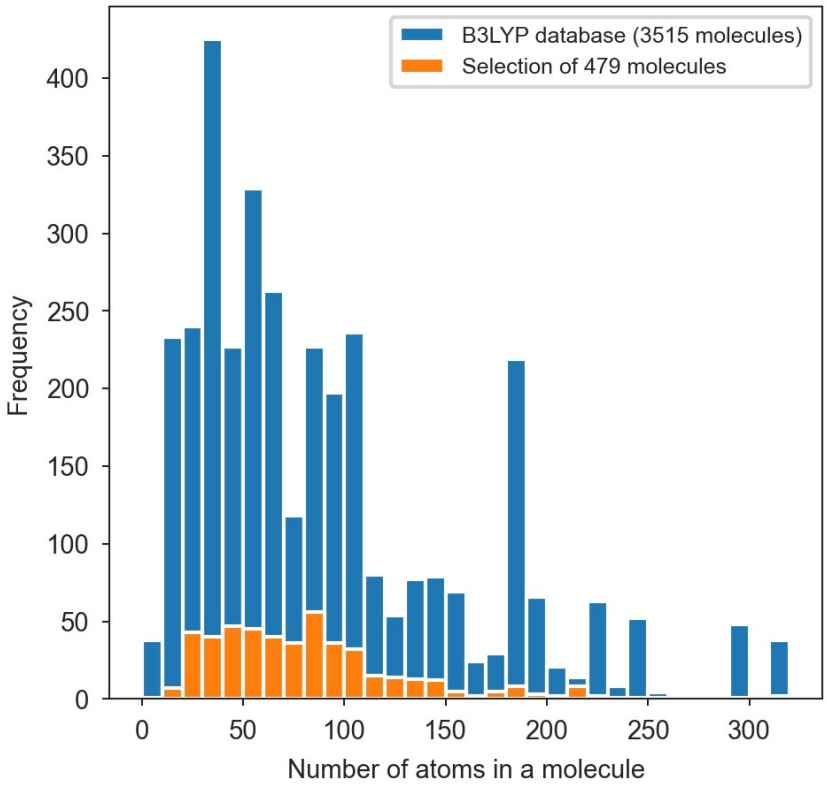


Figure S7. Histogram of the number of atoms present in the molecules of our TDDFT (B3LYP) dataset. Blue bars correspond to the molecules found originally in the dataset (3515 entries). Orange bars represent the distribution of the number of atoms found in the 479 molecules selected based on lowest energy conformation criteria.

Supplementary Note 2

Supplementary Note 2. Description of the TDDFT database, statistical and machine-learning methods.

The pristine data source of this work consists of a database of 3515 molecules optimized via DFT using the B3LYP functional as implemented in Gaussian09 software package. The database gathers original calculations performed for this particular study on conjugated small molecules as well as others developed in-house during the past years, including diverse conjugated small molecules, fullerenes and conjugated (co-)oligomers in distinct conformations (i.e., cis-/trans-).[13] Given the variety of input sources, the corresponding data cleaning procedure consists of: i) identifying duplicates based on molecular weight; and ii) picking the lowest energy molecular conformation among each set of duplicates. The filtering results in a final selection of 479 conjugated small molecules and oligomers optimized at the B3LYP level of theory. The resulting database gathers a variety of basis sets employed in the geometrical optimization step: 48% of the molecules were optimized using the 6-31G(d) set and 36% of them using the more computationally-expensive 6-311+G(d,p), see Figure S8. Furthermore, the chemical heterogeneity of the studied database is leaned toward known molecules and moieties of frequent use in high-performing solar energy harvesting applications, see Figure S9 and Figure S10. Side chains are systematically omitted or substituted by methyl groups in all calculations to reduce the computational cost.

Figure S8

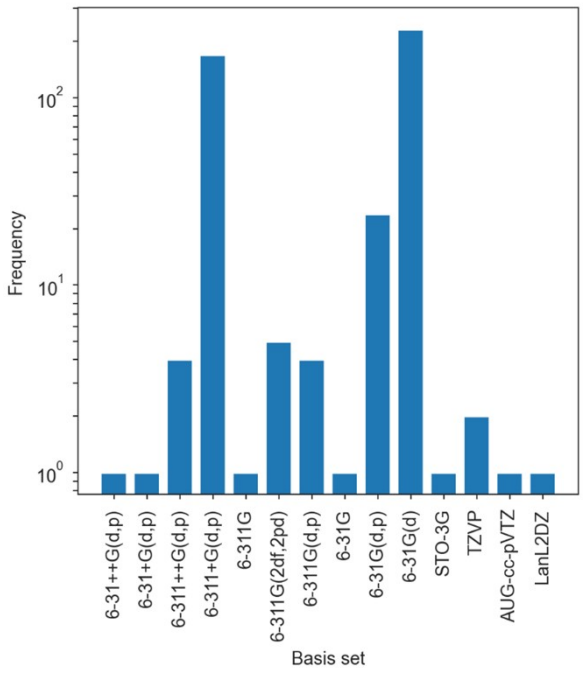


Figure S8. Histogram of the basis set employed in the geometrical optimization of the 479 molecules found in the curated DFT database.

Figure S9

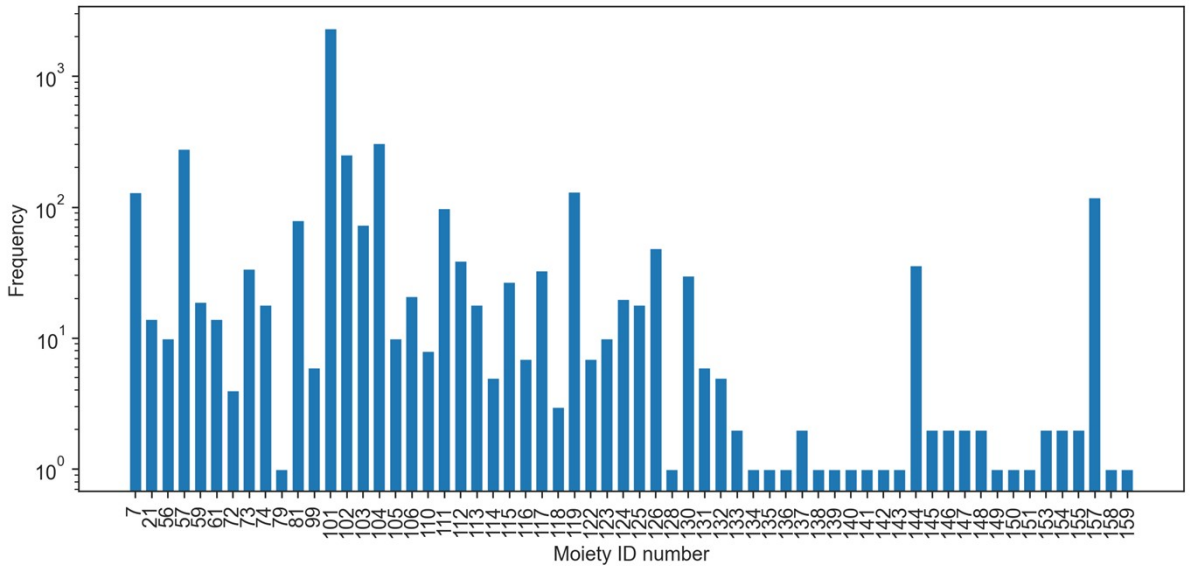


Figure S9. Histogram of the moieties present in the DFT database of 479 conjugated small molecules and oligomers. Moieties labels correspond to the chemical structures shown in Supplementary Note 4.

Figure S10

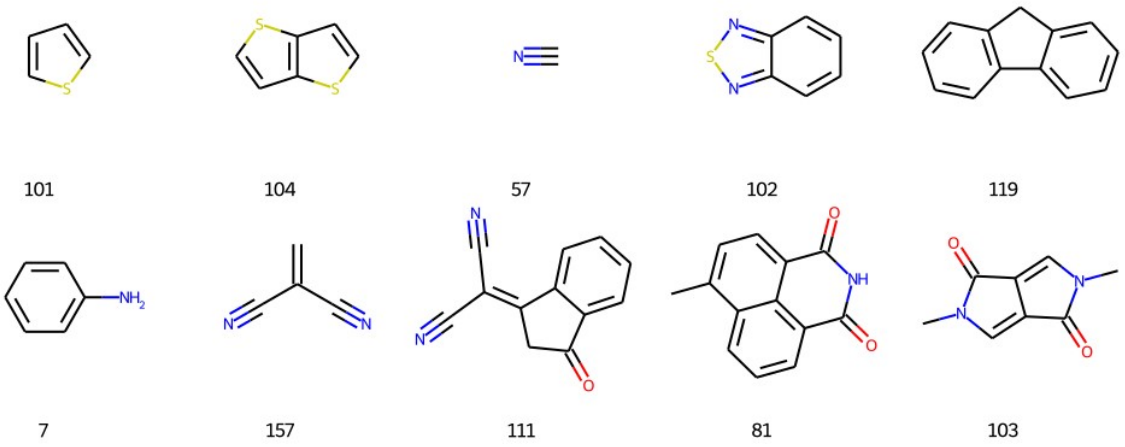


Figure S10. The 10 most frequent moieties together with their corresponding ID number.

Molecular descriptors are computed using four different open-source software packages and Python libraries (such as NumPy),[14] including 1D, 2D and 3D descriptors as retrieved from the corresponding DFT optimized geometries. The software bundle employed includes PaDEL (1874 descriptors),[15] PyChem (1094 descriptors),[16] Mordred (1826 descriptors)[17] and RDKit v2021.03.2. (1039 descriptors).[18] As a result, we obtained an initial set of 5834 descriptors for each molecule, which decreased up to 3239 after curing (i.e. dropping of uninformative or constant descriptors and others containing infinite or NaN values). Note that the same descriptor might be computed by more than one software bundle, yet slight numerical disagreements may arise due to the different computation algorithms. For that reason, we do not filter out redundant descriptors and perform their subsequent statistical analysis using the entire available catalogue. Furthermore, we include electronic features retrieved from the DFT calculations such as the energy levels of the 20 occupied and unoccupied molecular orbitals (HOMOs and LUMOs) closer to the band gap; the electronic band gap energy itself, E_{gap}; and the number of π electrons (n_π) in the molecule, which was determined using custom coding based on the RDKit library. The set of molecular fingerprints tested in this work is computed using RDKit and it includes customized coding for the moiety fingerprints and built-in functions for the computation of MACCS keys, Morgan fingerprints,[19] path-based (topological) fingerprints, E-state fingerprints[20] and Coulomb vectors.[21]

The target features in this study focus on the maximum oscillator strength (f_{max}) and other derived figures such as the maximum oscillator strength in the visible electromagnetic spectrum ($f_{max,vis}$, herein constrained between 300-1200 nm for its relevance in solar energy harvesting

applications); the sum of f in the visible window, $f_{max,vis}$; the spectral overlap between the Gaussian-broadened spectrum of fs in the visible (taking a standard deviation of 0.1 eV) and the AM1.5G solar irradiance spectrum, $f_{overlap}$; the maximum absorption coefficient (α_{max}); the maximum of the imaginary part of the dielectric function ($\varepsilon_{2,max}$); and the maximum molar extinction coefficient, f_{max}, $f_{max,vis}$ and $f_{sum,vis}$ are also evaluated per number of π electrons in the molecule, i.e. f_{max}/n_{π}, $f_{max,vis}/n_{\pi}$ and $f_{sum,vis}/n_{\pi}$.

The statistical analysis of descriptors is deployed using the open-source library SciPy[22] whereas the machine-learning (ML) models (k-nearest neighbours, linear regression, support vector regressor and random forests) are implemented in Scikit-Learn.[23]

Regarding the scoring of the ML models, R^2 ranges from -∞ to unity, being 1 the best possible score and zero an indication of lack of predictive power (as it is always returning the expected value of the target function, i.e., its average value); R^2_{adj} is formulated as[24]

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where p is the number of variables and n the sample size. Thus, R^2_{adj} adds penalties if the model uses too many variables, which is a useful metric when studying feature selection procedures. Test sets comprise 30% of the available data and all models are 10-fold holdout cross-validated (unless otherwise stated, using a randomized 70%-30% splitting for the train and test sets, respectively).

The recursive feature elimination (RFE) procedure applied in this work starts by decreasing the number of input features to 32 (i.e., 1% of the starting descriptor population of 3239 descriptors) in six consecutive feature reduction steps, in which after performing successive 10-fold cross-validations we drop 50% of the (averaged and least important) descriptors. Rather than observing a performance drop, the actual scoring of the RF ensemble improves as the number of features is reduced from 3239 ($R^2 = 0.65 \pm 0.06, r = 0.82 \pm 0.03$) to 51 ($R^2 = 0.70 \pm 0.05, r = 0.85 \pm 0.02$) variables in the last RFE iteration (**Figure S11**). After the last pruning step (51 variables), we select the 32 most important descriptors and perform a more thorough feature selection procedure by successively dropping (one-by-one) the least important descriptor (always keeping a 10-fold cross-validation scheme, see **Figure S12**).

Figure S11

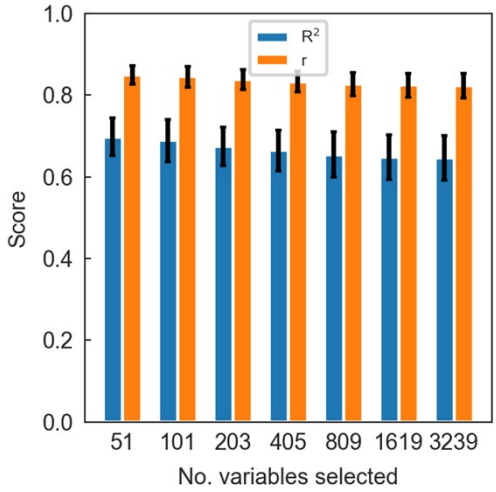


Figure S11. Scoring of RF regressors as part of a recursive feature elimination (RFE) loop in which 50% of the least important descriptors are dropped at each step.

Figure S12

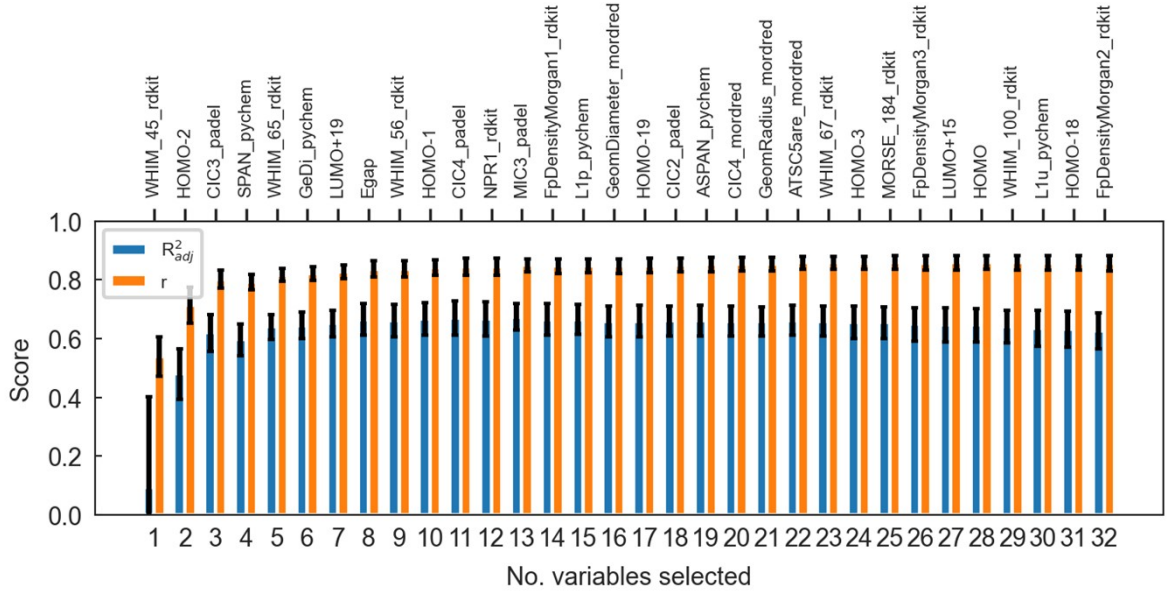


Figure S12. Scoring of 10-fold cross-validated RF regressors (300 estimators) trained and tested using different amounts of input descriptors as progressively indicated by the RFE algorithm. The top axis indicates, from left to right, the name of the variable that is added to the RF model, thus forming an ordered list of the most important descriptors found by the RF method.

Figure S13

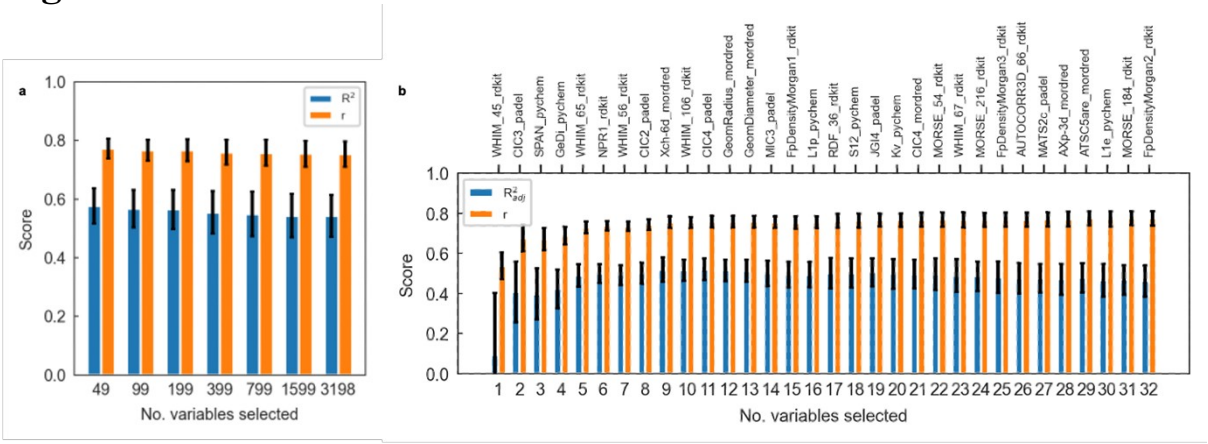


Figure S13. Performance of RF regressors trained without including electronic descriptors, using 300 estimators and 10-fold cross-validation. (a) Scoring parameters of cross-validated models as part of the RFE algorithm. (b) Detailed scoring parameters of the last 32 models obtained by RFE.

Supplementary Note 3

Supplementary Note 3. Clusterization algorithm of multicollinear descriptors.

The clusterization algorithm starts by taking the descriptor with the highest Spearman's rank correlation coefficient (ρ) and computing the Pearson correlation coefficient (r) with respect to the remaining elements in the ρ-ordered list of descriptors with $\rho \geq 0.7$. Descriptors from this list are dropped if $r \geq 0.7$ and considered to be in the same cluster; those showing $r \leq 0.7$ are candidates to form a different cluster. The process runs in a recursive-elimination manner until naturally leading to a selection of (typically) 1 to 5 descriptor clusters depending on the selected thresholds (0.6-0.7). These clusters gather the most statistically relevant and monotonic correlation trends with the target feature. Interestingly, by looking at the features stored in each of the clusters it is possible to replace some of the descriptors found originally by the algorithm by alternative figures of easier interpretation and/or larger physicochemical relevance.

Figure S14

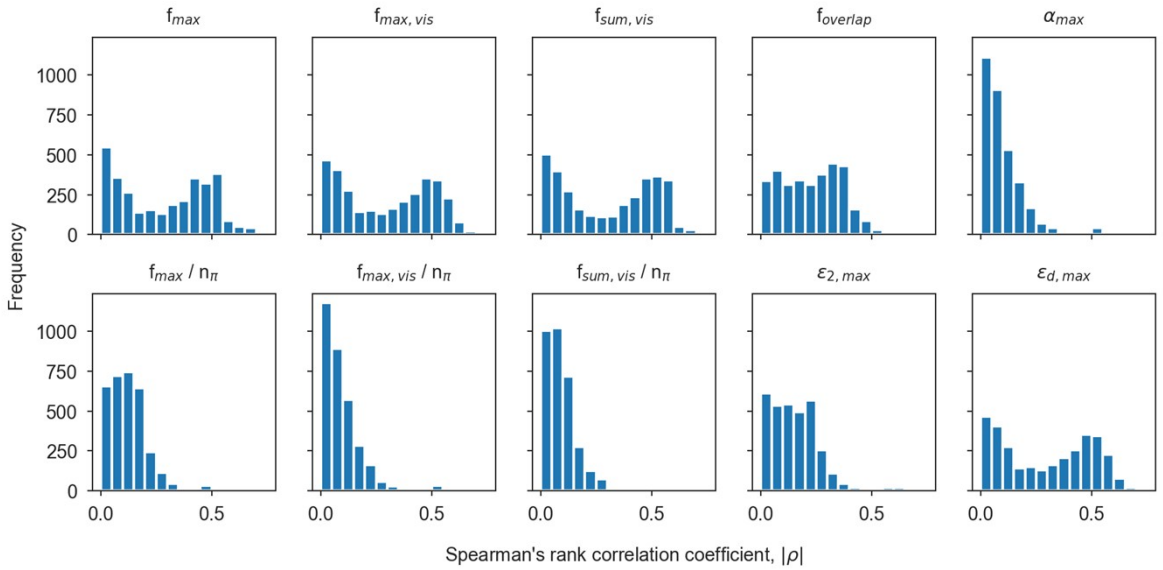


Figure S14. Spearman's rank correlation coefficient (in absolute value) histograms for the 3239 descriptors and the 10 different target features related with optical absorption and oscillator strength explored in this work.

Figure S15

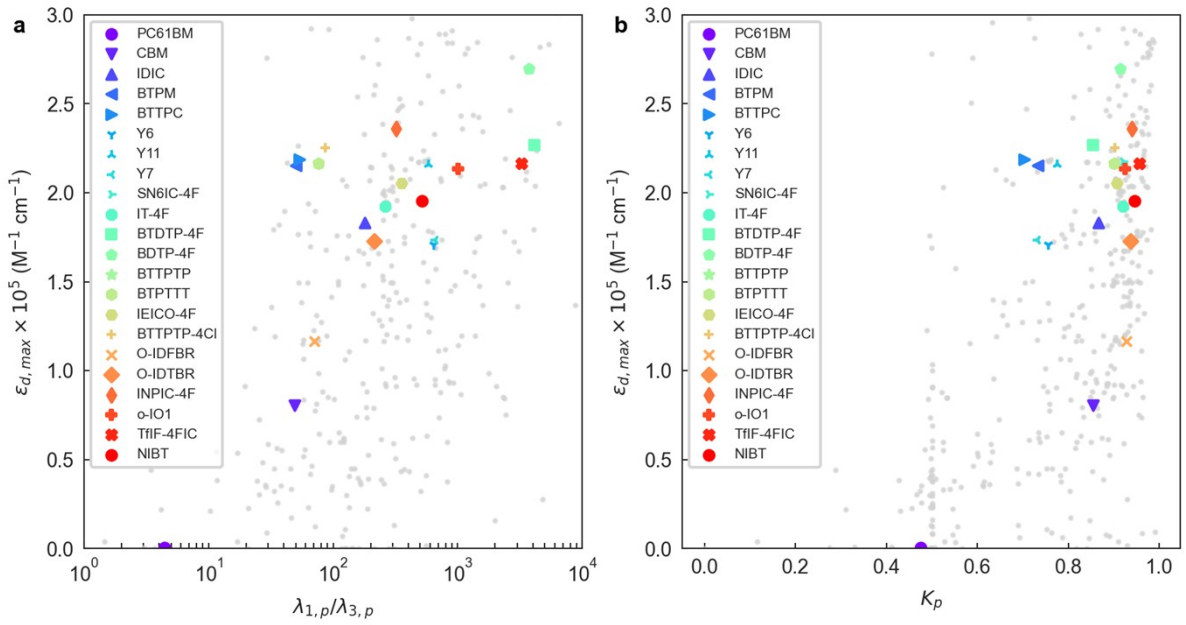


Figure S15. Influence of the molecular planarity on the maximum molar extinction coefficient. (a) The $\lambda_{1,p}/\lambda_{3,p}$ ratio correlates positively with $\varepsilon_{d,max}$ as straighter (more linear) molecules show larger $\lambda_{1,p}$ while enhanced molecular planarity lowers $\lambda_{3,p}$ values (which approach zero as there is no variance out of the molecular plane). (b) The shape of the molecule quantified with K_p shows that linear and planar molecules (i.e., K_p closer to unity)[25] enable larger $\varepsilon_{d,max}$ values. K_p is defined as[25]

$$K_p = \frac{\sum_m \left| \frac{\lambda_{m,p}}{\sum_m \lambda_{m,p}} - \frac{1}{3} \right|}{4/3},$$

where $m = 1,2,3$ and $0 \leq K_p \leq 1$.

Figure S16

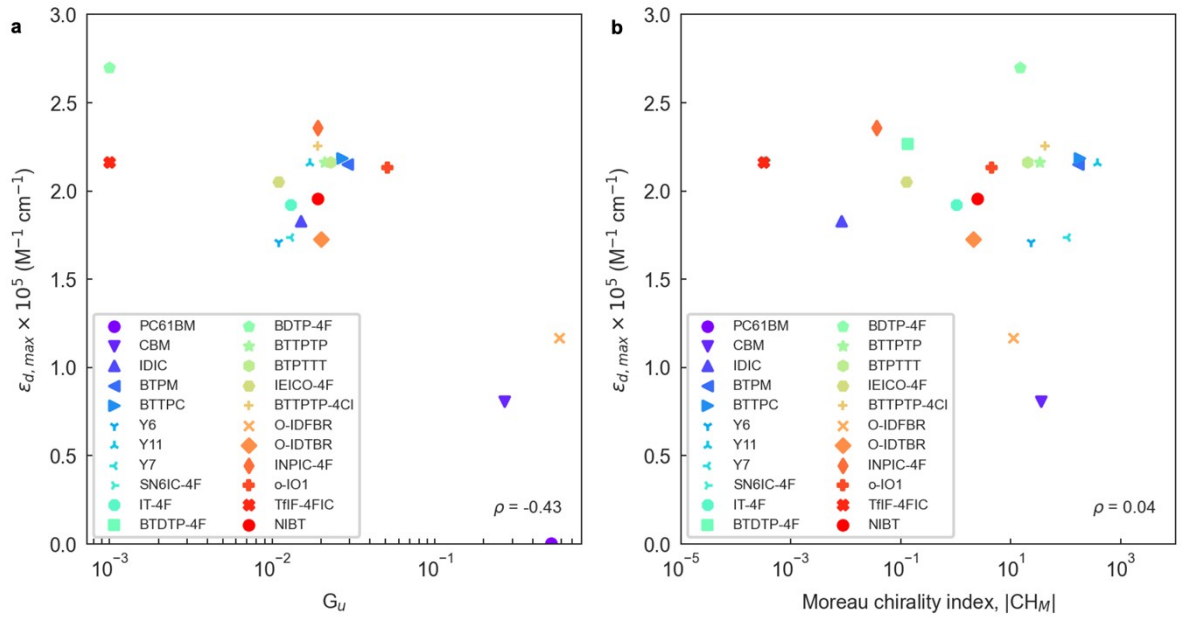


Figure S16. Influence of molecular symmetry on absorption strength. (a) Quantification of the total molecular symmetry as per the definition of the WHIM symmetry descriptor G_u (corresponding to the unweighted geometric mean of the directional symmetries, $G_u = \sqrt[3]{\gamma_{1,u} \cdot \gamma_{2,u} \cdot \gamma_{3,u}}$)[25] shows that as the molecules lose their central symmetry (i.e., lower G_u values), $\varepsilon_{d,max}$ can be further enhanced. (b) Conversely, the Moreau chirality index[26] weighted by atomic coordinates of small molecular absorbers shows poor correlation with $\varepsilon_{d,max}$.

Figure S17

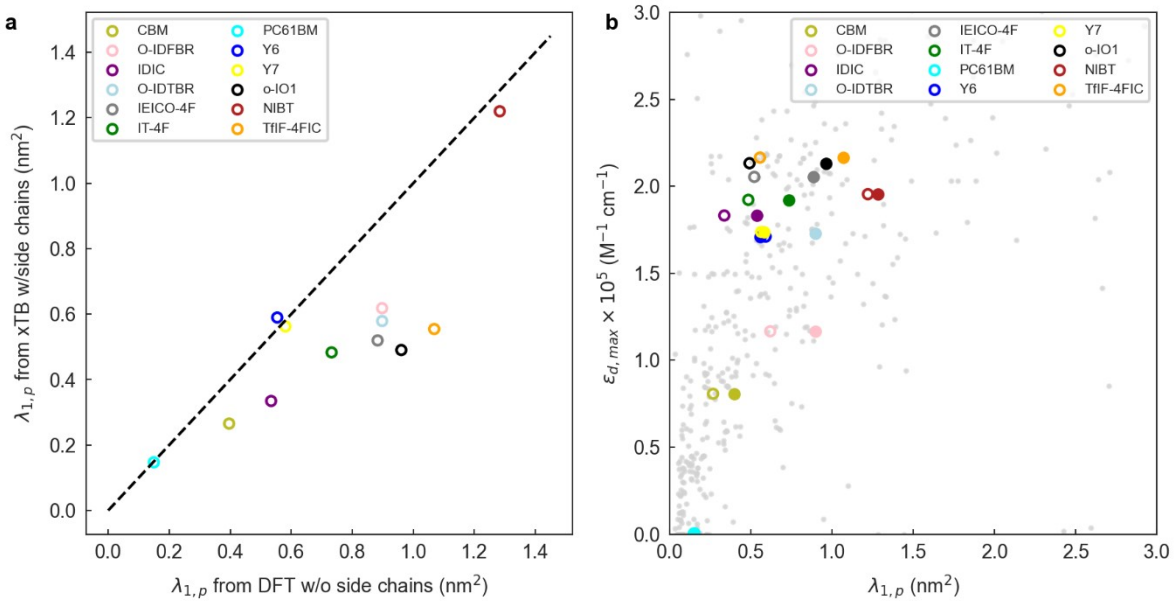


Figure S17. Influence of the side chains on $\lambda_{1,p}$ values. (a) Comparison of $\lambda_{1,p}$ values for a selection of small molecule acceptors as computed from xTB including side chains (y axis) and DFT-optimized geometries with methyl-substituted side chains (x axis). (b) Maximum molar extinction coefficient ($\varepsilon_{d,max}$) as a function of $\lambda_{1,p}$ for small molecule acceptors optimized with (open circles) and without (filled circles) side chains.

Figure S18

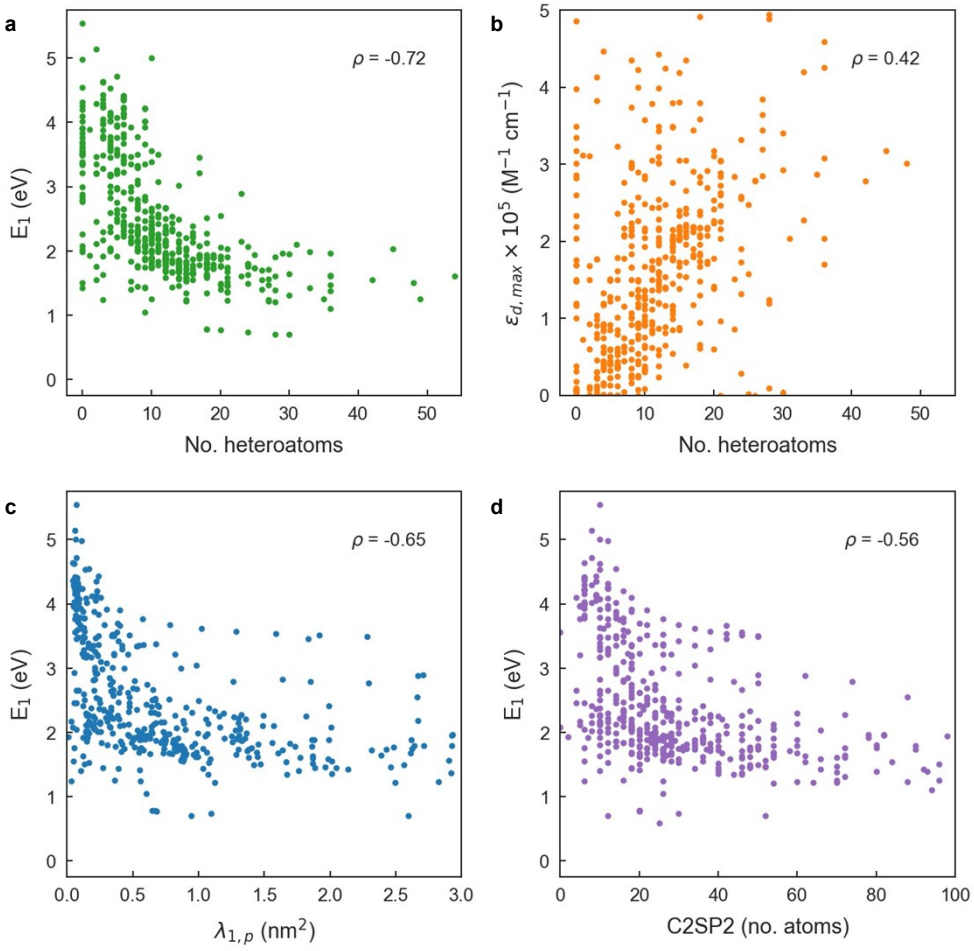


Figure S18. (a) Correlation between E_1 and the number of heteroatoms in the molecules. (b) Correlation between the molar extinction coefficient ($\varepsilon_{d,max}$) and the number of heteroatoms. (c) Correlation between E_1 and $\lambda_{1,p}$. (d) Correlation between E_1 and C2SP2. All panels include the corresponding Spearman's rank correlation coefficient (ρ).

Figure S19

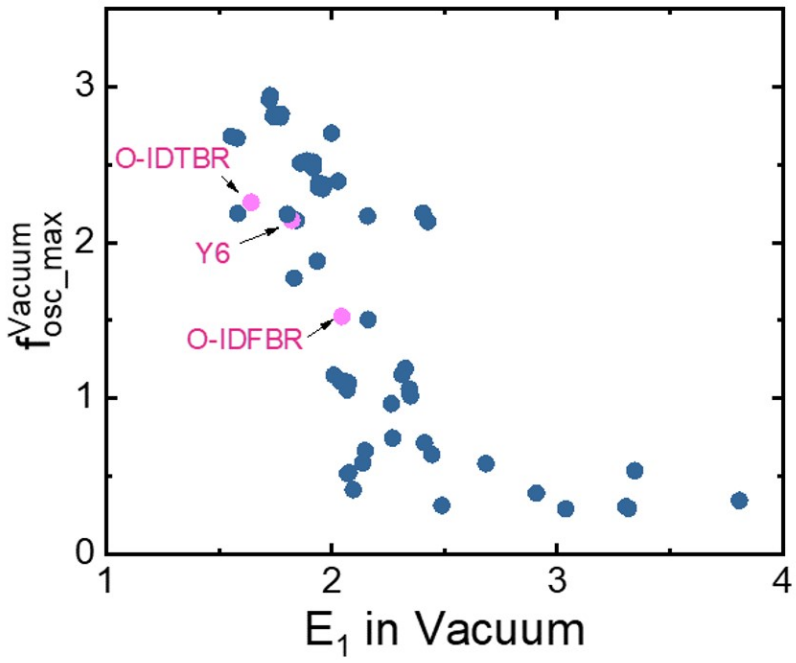


Figure S19. Relationship between the maximum oscillator strength and the energy of the first electronic transition in a set of TDDFT-optimized NFAs.

Table S1

Table S1. Statistical performance of a manifold of 10-fold cross-validated baseline models using $\varepsilon_{d,max}$ as target feature.

Model	**No. variables**	**R^2**	**r**
1-nearest neighbour	2	-0.18 ± 0.41	0.50 ± 0.05
	3239	-0.10 ± 0.37	0.47 ± 0.09
Linear regression	2	0.37 ± 0.10	0.61 ± 0.09
Random forest w/300 estimators	2	0.23 ± 0.17	0.59 ± 0.06
	3239	0.65 ± 0.06	0.82 ± 0.03

Figure S20

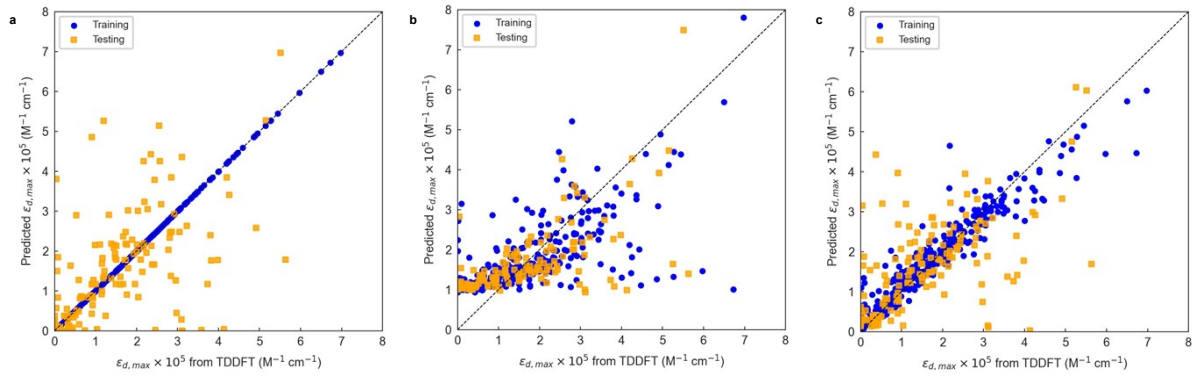


Figure S20. Correlation plots of three (exemplary) baseline models trained and tested on $\varepsilon_{d,max}$ using two input descriptors only: $\lambda_{1,p}$ and C2SP2. (a) 1-nearest neighbour; (b) linear regression; and (c) out-of-the-box RF trained with 300 estimators.

Table S2

Table S2. Performance of RF models trained and 10-fold cross-validated using 300 estimators, 3 input molecular descriptors ($\lambda_{1,v}$, CIC3 and HOMO-2) and different forms of molecular fingerprint vectors. In the case of Morgan fingerprints, we set the connectivity radius to 4 units, while for topological fingerprints the minimum and maximum path counts are set to 1 and 6 units, respectively. Their vector lengths are set to either 64 or 2048 bits to reflect different degrees of model complexity.

No. molecular descriptors	Fingerprint type	No. bits	Total no. inputs	R^2	r
3	N/A	N/A	3	0.63 ± 0.06	0.80 ± 0.03
3	Moiety	159	162	0.63 ± 0.06	0.81 ± 0.04
3	MACCS	166	169	0.66 ± 0.04	0.83 ± 0.02
3	Morgan	64	67	0.70 ± 0.05	0.84 ± 0.03
		2048	2051	0.69 ± 0.05	0.84 ± 0.03
3	Topology	64	67	0.68 ± 0.04	0.83 ± 0.02
		2048	2051	0.69 ± 0.05	0.84 ± 0.03
3	E-state	79	82	0.66 ± 0.04	0.82 ± 0.02
3	Coulomb	320	323	0.56 ± 0.07	0.77 ± 0.04

Table S3

Table S3. Scoring of the baseline and hyperparametrically optimized RF and ExtraTrees models, fed with 3 molecular descriptors and a Morgan fingerprint vector of 64 bits.

Model	No. estimators	No. samples per leaf	No. samples to split	Validation	R^2	r
RF (out-of-the-box)	300	1	2	10-fold CV	0.70 ± 0.05	0.84 ± 0.03
RF (optimized)	1200	1	2	10-fold CV	0.70 ± 0.05	0.85 ± 0.03
RF (optimized)	1200	1	2	LOOCV	0.74	0.86
ExtraTrees (out-of-the-box)	300	1	2	10-fold CV	0.69 ± 0.05	0.85 ± 0.02
ExtraTrees (optimized)	2000	1	2	10-fold CV	0.70 ± 0.04	0.85 ± 0.02
ExtraTrees (optimized)	2000	1	2	LOOCV	0.73	0.86

Figure S21

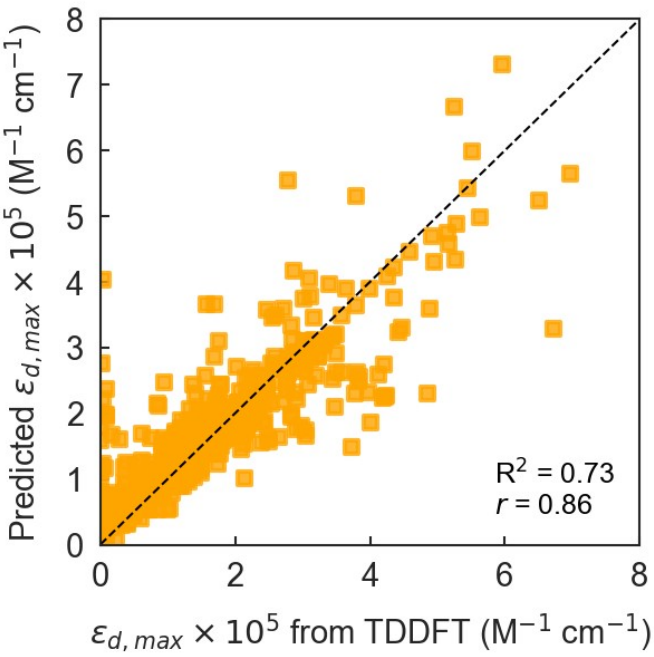


Figure S21. LOOCV of the optimized Extra Trees (ET) regressor fed with 3 molecular descriptors and a 64-bit vector as Morgan fingerprint.

Figure S22

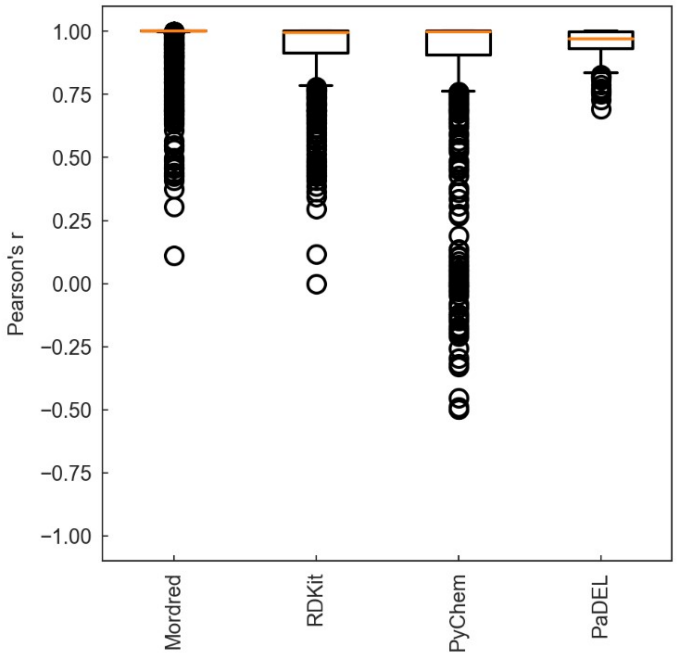


Figure S22. Boxplots for the Pearson correlation coefficients between different sets of molecular descriptors retrieved from xTB and DFT (B3LYP) optimized geometries.

Figure S23

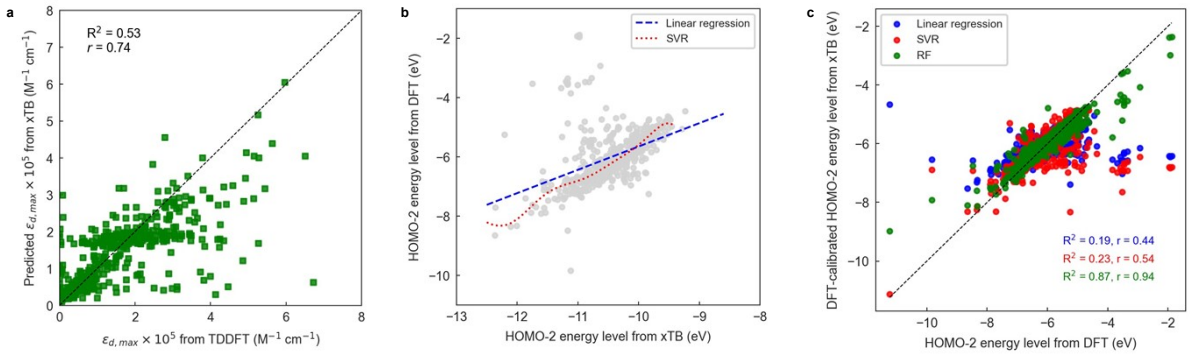
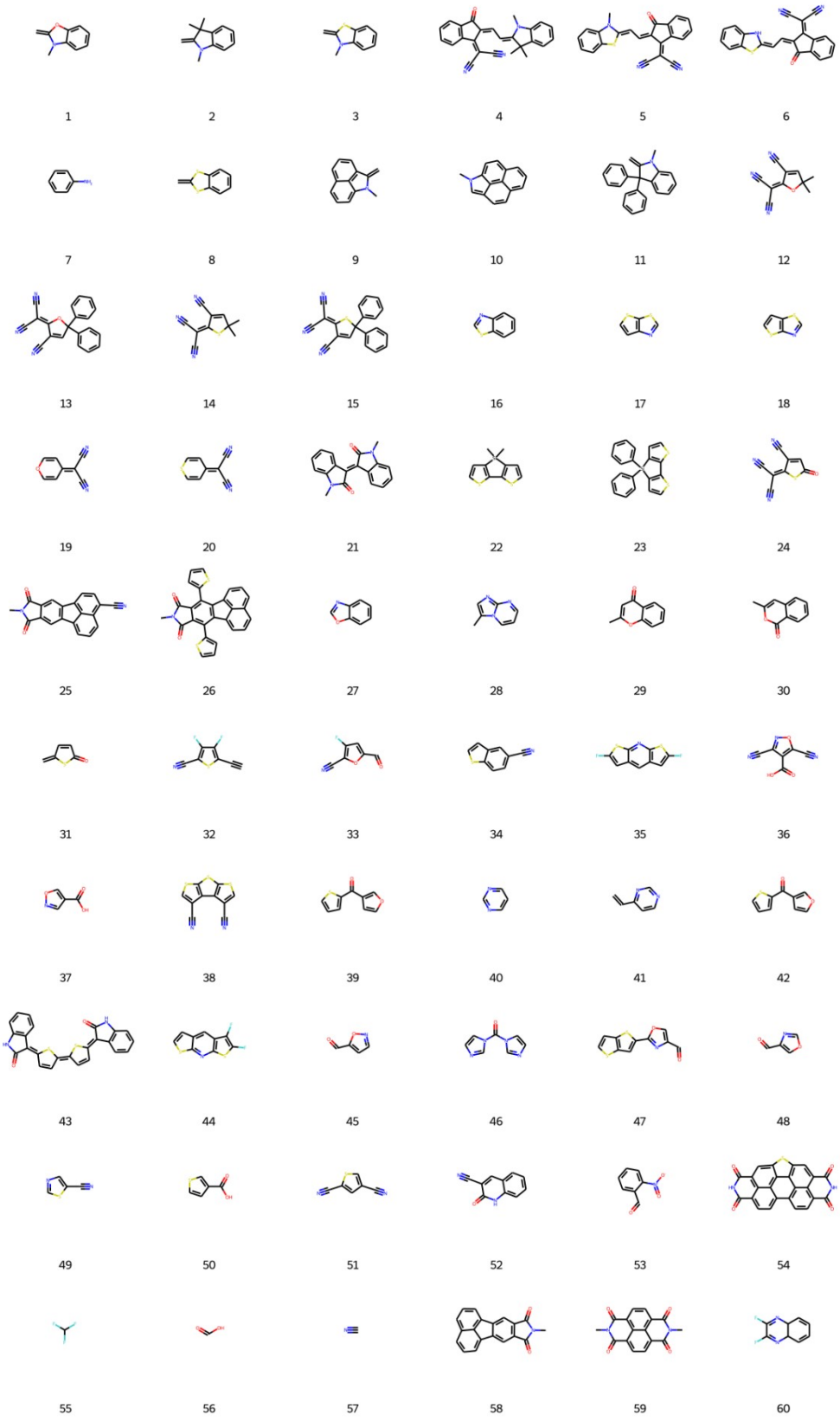
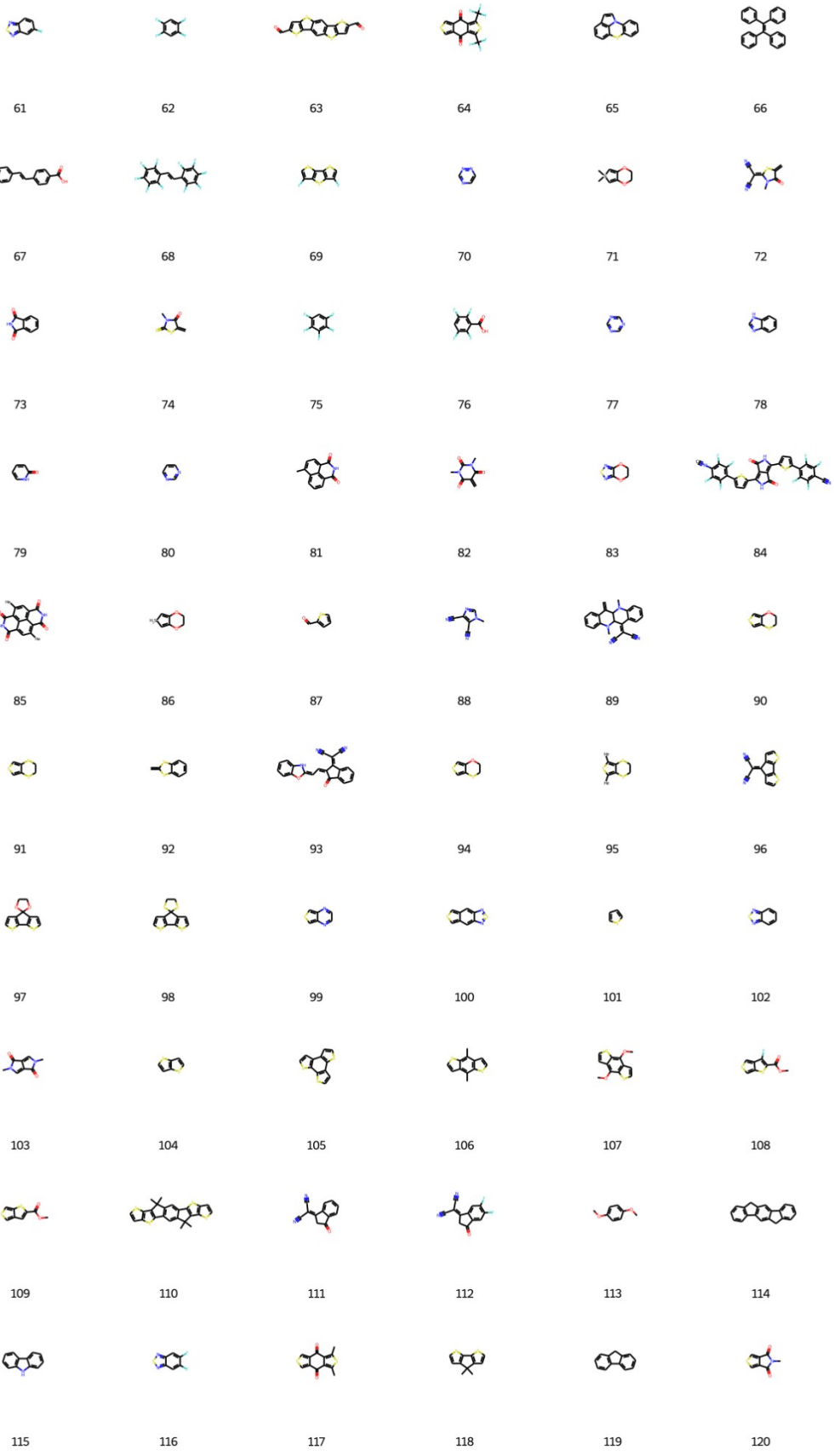


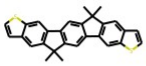
Figure S23. (a) Leave-one-out interpolation of a RF model trained using DFT data and tested on xTB-optimized molecules using 3 parameters ($\lambda_{1,v}$ and CIC3 recalculated from the xTB geometry, and the HOMO-2 energy level as computed in xTB) and a 64-bit vector as Morgan fingerprint. (b) Fitting of linear regression and support vector regressor (SVR) models for calibration of HOMO-2 energy levels as computed in xTB and DFT (B3LYP). The mismatch in the absolute values of HOMO-2 energy levels between DFT and xTB calculations prevents obtaining higher scorings in the RF models depicted in (a). (c) Correlation plot between HOMO-2 energy levels from DFT and the corresponding calibrated values as obtained by linear regression (blue), SVR (red) and RF (green) models. The dashed black line indicates perfect matching between DFT and calibrated values.

Supplementary Note 4

Supplementary Note 4. Detailed database of moieties used in this work.

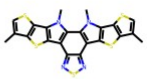



61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120



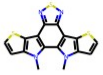
121

122

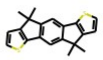


123

124



125



126

127

128

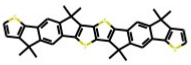
129

130

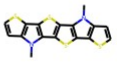
131

132

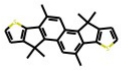
133



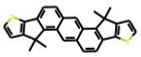
134



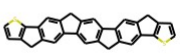
135



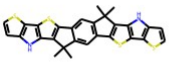
136



137



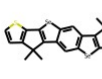
138



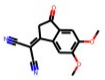
139

140

141



142



143

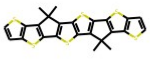
144

145

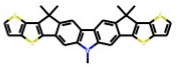
146

147

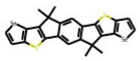
148



149



150



151

152

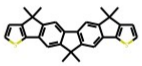
153

154

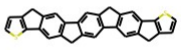
155

156

157



158



159

Supplementary Note 5

Supplementary Note 5. Estimation of computation time required to make absorption strength predictions using xTB Hamiltonians (in combination with ML models) or rigorous TDDFT.

Table S4 provides a comparison in terms of the computation time required for the molecular geometry optimization step in the TDDFT and xTB approaches. We analysed 194 molecules from TDDFT calculations based on B3LYP/6-311+G(d,p) and 475 molecules (**Figure S7**) from xTB calculations based on GFN2–xTB to tentatively quantify the difference in computational efficiency between both methods. Our analysis suggests that geometry optimization using GFN2–xTB is ca. 3000 times faster than TDDFT/B3LYP/6-311+G(d,p), even though GFN2–xTB calculations were done on a conventional 12 CPUs laptop as opposed to the 32 CPUs dedicated cluster/workstation employed in the TDDFT calculations, thus highlighting the great advantage of using xTB over TDDFT.

Furthermore, we have estimated the time consumption for the absorption strength predictions using the established ML model in this work. The time required for the ML model training and LOOCV steps is below 10 minutes (12 CPUs), whereas the calculation of molecular descriptors (>5000 descriptors) for the full data set (479 molecules) takes no less than 180 minutes (12 CPUs). Hence, for a molecule made up of 100 atoms, the whole absorption strength determination (i.e., from geometry optimization to $\varepsilon_{d,max}$ prediction) effectively takes around 200 minutes using xTB with ML; and 1345 minutes using solely TDDFT. Nevertheless, the advantage of the ML approach is more evident as interpolation in the trained model takes less than 1 second (per molecule) to compute, which enables at least four orders of magnitude faster molecular screening with respect to TDDFT (1345 minutes or 80700 seconds per molecule).

Table S4

Table S4. Computation time required for molecular geometry optimization steps using TDDFT/B3LYP/6-311+G(d,p) and xTB/GFN2–xTB.

Approach	DFT/B3LYP/6-311+G(d,p)	xTB/GFN2–xTB
No. of molecules	194	475
No. of atoms	18022	37989
No. of CPUs	32	12
Time elapsed (mins)	206398.355	136
Time elapsed per atom (mins)	11.45257768	0.003579984

References

1 S. Karuthedath, J. Gorenflot, Y. Firdaus, N. Chaturvedi, C. S. P. De Castro, G. T. Harrison, J. I. Khan, A. Markina, A. H. Balawi, T. A. Dela Peña, W. Liu, R.-Z. Liang, A. Sharma, S. H. K. Paleti, W. Zhang, Y. Lin, E. Alarousu, D. H. Anjum, P. M. Beaujuge, S. De Wolf, I. McCulloch, T. D. Anthopoulos, D. Baran, D. Andrienko and F. Laquai, *Nat. Mater.*, 2021, **20**, 378–384.

2 W. Wang, B. Zhao, Z. Cong, Y. Xie, H. Wu, Q. Liang, S. Liu, F. Liu, C. Gao, H. Wu and Y. Cao, *ACS Energy Lett.*, 2018, **3**, 1499–1507.

3 S. Holliday, R. S. Ashraf, A. Wadsworth, D. Baran, S. A. Yousaf, C. B. Nielsen, C. H. Tan, S. D. Dimitrov, Z. Shang, N. Gasparini, M. Alamoudi, F. Laquai, C. J. Brabec, A. Salleo, J. R. Durrant and I. McCulloch, *Nat. Commun.*, 2016, **7**, 1–11.

4 D. Baran, T. Kirchartz, S. Wheeler, S. Dimitrov, M. Abdelsamie, J. Gorman, R. S. Ashraf, S. Holliday, A. Wadsworth, N. Gasparini, P. Kaienburg, H. Yan, A. Amassian, C. J. Brabec, J. R. Durrant and I. McCulloch, *Energy Environ. Sci.*, 2016, **9**, 3783–3793.

5 N. A. Cooling, E. F. Barnes, F. Almyahi, K. Feron, M. F. Al-Mudhaffer, A. Al-Ahmad, B. Vaughan, T. R. Andersen, M. J. Griffith, A. S. Hart, A. G. Lyons, W. J. Belcher and P. C. Dastoor, *J. Mater. Chem. A*, 2016, **4**, 10274–10281.

6 M. M. Wienk, J. M. Kroon, W. J. H. Verhees, J. Knol, J. C. Hummelen, P. A. van Hal and R. A. J. Janssen, *Angew. Chemie Int. Ed.*, 2003, **42**, 3371–3375.

7 D. Baran, R. S. Ashraf, D. A. Hanifi, M. Abdelsamie, N. Gasparini, J. A. Röhr, S. Holliday, A. Wadsworth, S. Lockett, M. Neophytou, C. J. M. Emmott, J. Nelson, C. J. Brabec, A. Amassian, A. Salleo, T. Kirchartz, J. R. Durrant and I. McCulloch, *Nat. Mater.*, 2017, **16**, 363–369.

8 M. Li, Y. Liu, W. Ni, F. Liu, H. Feng, Y. Zhang, T. Liu, H. Zhang, X. Wan, B. Kan, Q. Zhang, T. P. Russell and Y. Chen, *J. Mater. Chem. A*, 2016, **4**, 10409–10413.

9 N. Qiu, H. Zhang, X. Wan, C. Li, X. Ke, H. Feng, B. Kan, H. Zhang, Q. Zhang, Y. Lu and Y. Chen, *Adv. Mater.*, 2017, **29**, 1604964.

10 Z. Cong, B. Zhao, Z. Chen, W. Wang, H. Wu, J. Liu, J. Wang, L. Wang, W. Ma and C. Gao, *ACS Appl. Mater. Interfaces*, 2019, **11**, 16795–16803.

11 W. Zhao, S. Li, H. Yao, S. Zhang, Y. Zhang, B. Yang and J. Hou, *J. Am. Chem. Soc.*, 2017, **139**, 7148–7151.

12 G. Forti, A. Nitti, P. Osw, G. Bianchi, R. Po and D. Pasini, *Int. J. Mol. Sci.*, 2020, **21**, 8085.

13 M. S. Vezie, S. Few, I. Meager, G. Pieridou, B. Dörling, R. S. Ashraf, A. R. Goñi, H. Bronstein, I. McCulloch, S. C. Hayes, M. Campoy-Quiles and J. Nelson, *Nat. Mater.*, 2016, **15**, 746–753.

14 C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.

15 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.

16 D.-S. Cao, Q.-S. Xu, Q.-N. Hu and Y.-Z. Liang, *Bioinformatics*, 2013, **29**, 1092–1094.

17 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminform.*, 2018, **10**, 4.

18 RDKit: Open-Source Cheminformatics Software.

19 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

20 L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1039–1045.

21 D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge and P. W. Chung, *Sci. Rep.*, 2018, **8**, 9059.

22 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa and P. van Mulbregt, *Nat. Methods*, 2020, **17**, 261–272.

23 F. Pedregosa, R. Weiss, M. Brucher, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

24 N. S. Raju, R. Bilgic, J. E. Edwards and P. F. Fleer, *Appl. Psychol. Meas.*, 1997, **21**, 291–305.

25 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley, Second Edi., 2009, vol. 41.

26 G. Moreau, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 929–938.