SUPPORTING INFORMATION

to **Evidence conditioned screening prioritizes donor acceptor materials**

Jan Zdražil[1,2†], Tomáš Hrivnák[3,4†], Miroslav Medveď[1,5], Václav Snášel[2], Michal Otyepka[1,3]*

[1] Czech Advanced Technology and Research Institute (CATRIN), Regional Centre of Advanced Technologies and Materials, Palacký University, Šlechtitelů 27, Olomouc, 779 00 Czech Republic

[2] Faculty of Electrical Engineering and Computer Science, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava-Poruba 70800, Czech Republic

[3] IT4Innovations, VSB – Technical University of Ostrava, 17. listopadu 2172/15, Ostrava-Poruba 70800, Czech Republic

[4] Department of Molecular Simulation of Polymers, Polymer Institute, Slovak Academy of Sciences, Dúbravská cesta 9, 845 41 Bratislava, Slovakia

[5] Department of Chemistry, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 974 01 Banská Bystrica, Slovakia

[†] These authors contributed equally to this work.

* Corresponding author: michal.otyepka@upol.cz

Computational details – Dataset preparation

Monomer input dataset

The library of 76 molecular monomers utilized in this study was identical to that reported in our previous work.[1] In general, the library included medium-size polycyclic aromatic hydrocarbons (anthracene, pyrene, and perylene) and their nitrogen-doped derivatives functionalized by various electron-donating or electron-withdrawing groups at the synthetically relevant positions. Each monomer was represented using three complementary descriptor sets capturing molecular-level, atom-wise, and bond-wise properties, with the latter two partially inspired by Jia et al.[2]

At the molecular level, 40 descriptors were considered, combining both integer- and float-valued quantities. The complete list of descriptors together with their feature families used for XAI analysis is provided in Table 1. Based on the methodology used for feature generation, the descriptors can be divided into three categories:

- descriptors derived directly from the initial SMILES-generated structures, primarily using RDKit,[3] or from RDKit-generated geometries (radius of gyration, asphericity, and prolateness);
- descriptors evaluated at the ωB97XD/def2-TZVP level of theory[4,5] in vacuum following geometry optimization employing a smaller def2-SVP basis set;
- descriptors analogous to category (2), but evaluated for dichloromethane (DCM) solvent environment using the SMD solvation model[6] with a solvent-excluded surface.

All quantum-mechanical (QM) calculations were performed using Gaussian16,[7] except for excited state computations which were performed at the ωB97X-D3/def2-TZVP level of theory[4,8] using simplified Tamm-Dancoff (sTDA) approach[9] in ORCA 5.0.3.[10,11] The bright state was defined as the lowest-energy excited state with an oscillator strength greater than 0.1. The radius of gyration (R_g), asphericity (Δ), and prolateness (S) were evaluated from the eigenvalues of the moment of inertia tensor (I) according to following relations:

$$R_g = \sqrt{\frac{\sum_i {\lambda_i}^2}{2}}$$

$$\Delta = 1 - 3\sum_{i\neq j}\left(\frac{\lambda_i}{R_g}\right)^2\left(\frac{\lambda_j}{R_g}\right)^2$$

$$S = \prod_i\left[3\left(\frac{\lambda_i}{R_g}\right)^2 - 1\right]$$

$$\lambda_i = \sqrt{I_i/M}$$

where $i = 1, 2, 3$ and M is the molecular mass.

Atom-wise descriptors were generated analogously to the molecular descriptors, with the addition of selected MMFF96 parameters[12–14] following automatic RDKit atom-type attribution. These descriptors are summarized in Table 2. Quantum-mechanical partial charges were obtained using the CHelpG scheme[15] at the ωB97XD/def2-TZVP level of theory using geometries optimized at the ωB97XD/def2-SVP level. Bond-wise descriptors were generated in a similar manner and are summarized in Table 3.

Table 1. Molecular-level descriptors of monomers included in the dataset. Descriptor families are indicated by superscript letters and defined in the table footnote. Here, S_1 and S_{bright} denote the $S_0 \rightarrow S_1$ and $S_0 \rightarrow S_{bright}$ electronic transitions, respectively.

SMILES-based features (1–11)	QM vacuum properties (12–25)	QM solute properties (26–40)
Number of atoms[a]	Radius of gyration[c] [Å]	Radius of gyration[e] [Å]
Number of bonds[a]	Asphericity[c]	Asphericity[e]
Molecular weight[a] [g/mol]	Prolateness[c]	Prolateness[e]
Logarithm of the partition coefficient[a]	E_{HOMO}[c] [eV]	E_{HOMO}[e] [eV]
Number of rotatable bonds[a]	E_{LUMO}[c] [eV]	E_{LUMO}[e] [eV]
Number of H-bond donors[a]	IP[c] [eV]	IP[e] [eV]
Number of H-bond acceptors[a]	EA[c] [eV]	EA[e] [eV]
Number of rings within molecule[a]	S_1 vertical excitation energy[d] [eV]	Solvation energy[e] [kcal/mol]
Radius of gyration[b] [Å]	S_1 oscillator strength[d] [au]	S_1 vertical excitation energy[f] [eV]
Asphericity[b]	S_1 HOMO→LUMO contribution[d]	S_1 oscillator strength[f] [au]
Prolateness[b]	S_{bright} vertical excitation energy[d] [eV]	S_1 HOMO→LUMO contribution[f]
	S_{bright} oscillator strength[d] [au]	S_{bright} vertical excitation energy[f] [eV]
	S_{bright} state number[d]	S_{bright} oscillator strength[f] [au]
	S_{bright} HOMO→LUMO contribution[d]	S_{bright} state number[f]
		S_{bright} HOMO→LUMO contribution[f]

[a] Counts/topology; [b] Global shape; [c] Orbital/energy (vacuum); [d] Excited state (vacuum); [e] Orbital/energy (solvent); [f] Excited state (solvent)

Table 2. Atom-level descriptors of monomers included in the dataset. Descriptor families are indicated by superscript letters and defined in the table footnote.

SMILES-based features (1–5)	MMFF96 features[b] (6–11)	QM vacuum properties[c] (12–13)
Atom number[a]	Atomic polarizability α_I [Å^3]	Partial charge [au]
Number of directly bonded neighbors[a]	Effective-electron number N_I	Distance from center of mass [nm]
Aromaticity[a]	A_I scale factor	QM solute properties[d] (14–15)
Hybridization[a]	G_I scale factor	Partial charge [au]
Distance from center of mass[b] [nm]	Minimum-vdW-energy separation R_{II} [Å]	Distance from center of mass [nm]
	Partial charge [au]	

[a] Atom identity/valence; [b] Atom geometry/MMFF; [c] Atom QM (vacuum); [d] Atom QM (solvent)

Table 3. Bond-level descriptors of monomers included in the dataset. Descriptor families are indicated by superscript letters and defined in the table footnote.

SMILES-based features[a] (1–5)	MMFF96 features[b] (6–7)	QM vacuum properties[c] (8)
Bond type	Bond stretching force constant [md/Å]	Optimized bond length [Å]
Is aromatic	Equilibrium bond length [Å]	QM solute properties[c] (9)
Is conjugated		Optimized bond length [Å]
Is in ring		
Bond order		

[a] Bond topology; [b] Bond MMFF; [c] Bond QM

Aggregate output dataset

Based on the criterion $E_{\mathrm{HOMO}}(\boldsymbol{B}) < E_{\mathrm{HOMO}}(\boldsymbol{A}) < E_{\mathrm{LUMO}}(\boldsymbol{B}) < E_{\mathrm{LUMO}}(\boldsymbol{A})$ evaluated in vacuum at the ωB97XD/def2-SVP level, 2445 unique donor–acceptor (D–A) molecular pairs were selected for analysis. Five pairs were subsequently excluded due to convergence issues of output descriptors.

For each pair, the molecular geometry was optimized using a three-step procedure. First, an initial conformational search was performed using an in-house code. The resulting structures were subsequently used as input for the CREST workflow.[16] The lowest-energy conformer obtained from CREST was then subjected to DFTB optimization in dichloromethane (DCM) solvent using DFTB+.[17] The DFTB3 OB3 parameter set[18,19] together with the COSMO solvation model including solvent accessible surface area contributions was utilized. The optimized geometries of the neutral systems were verified by frequency analysis to confirm the absence of imaginary frequencies and subsequently served as starting structures for optimization of the ionized systems.

The selected output descriptors are summarized in Table 4. The first eight descriptors can be divided into two groups of four associated with the $S_0 \rightarrow S_1$ vertical excitation and the bright excited state, respectively. The bright state was defined as the lowest-energy excited state with the oscillator strength greater than 0.1. If no such state was identified below 4.5 eV, the oscillator-strength threshold was iteratively reduced by a factor of 0.75 until a qualifying state was found. In addition to the excitation energy and oscillator strength, descriptors characterizing the nature of the electronic transition were also evaluated. These transitions were classified as predominantly local excitation (LE) or charge-transfer (CT) in character using fragment-based analysis of correlated electron-hole pairs, as introduced by Plasser et al.[20–22] and implemented in the TheoDORE program.[23] The corresponding omega matrices (Ω) were used to classify the transitions as predominantly: (1) A→A (LE), (2) A→B (CT), (3) B→A (CT), or (4) B→B (LE), together with the associated value of $\Omega^{OI}_{dominant}$. The omega matrices were normalized according to $\Omega^{0I} = \sum_{\boldsymbol{A},\boldsymbol{B}} \Omega^{OI}_{\boldsymbol{AB}} = 1$. Excited-state properties were evaluated at the sTDA-ωB97X-D3/def2-TZVP level of theory using the SMD solvation model with solvent excluded surface for DCM in ORCA 5.0.3. Only singlet excited states were considered.

The following three binary descriptors (0/1) characterize donor and acceptor aggregation preferences. These descriptors were determined by comparing Boltzmann populations at 298.15 K for the following aggregation scenarios:

- Formation of either separate **AA** and **BB** aggregates (0) or two **AB** aggregates (1);
- Formation of two **AB** aggregates and an isolated **A** monomer (0) versus formation of one **AB** aggregate, one **AA** aggregate, and an isolated **B** monomer (1);
- Formation of two **AB** aggregates and an isolated **B** monomer (0) versus formation of one **AB** aggregate, one **BB** aggregate, and an isolated **A** monomer (1).

The energies of all aggregates and monomers were evaluated at the ωB97XD/def2-SVP/SMD(DCM) level of theory with solvent excluded surface using Gaussian16 on DFTB-optimized geometries. Gibbs free energies additionally included vibrational corrections evaluated at the DFTB level.

Finally, the last two regression targets correspond to the adiabatic electron affinity and ionization potential of the D–A aggregate, excluding vibrational contributions.

Table 4. Output descriptors of D–A aggregates included in the dataset.

Name	Abbreviation	Description	Prediction type
S1_Excitation_Energy	S1 E	$S_0 \rightarrow S_1$ vertical excitation energy [eV]	Regression
S1_Oscillator_Strength	S1 f	$S_0 \rightarrow S_1$ oscillator strength	Regression
S1_Character	S1 char	$S_0 \rightarrow S_1$ dominant character class	Classification
S1_Dominant_Omega	S1 Ω	Ω of the $S_0 \rightarrow S_1$ character class	Regression
BS_Excitation_Energy	BS E	Bright state vertical excitation energy [eV]	Regression
BS_Oscillator_Strength	BS f	Bright state oscillator strength	Regression
BS_Character	BS char	S_{bright} dominant character class	Classification
BS_Dominant_Omega	BS Ω	Ω of the bright state character class	Regression
AB+AB_AA+BB	AB Mix	Preference for Donor-Acceptor stacking	Binary
AB+AB+A_AB+AA+B	A transfer	Preference for Donor-rich stacking	Binary
AB+AB+B_AB+BB+A	B transfer	Preference for Acceptor-rich stacking	Binary
EA	EA	Electron affinity [eV]	Regression
IP	IP	Ionization potential [eV]	Regression

References

1. Hrivnák, T., Pykal, M., Rogach, A. L., Otyepka, M. & Medveď, M. Computational design of donor–acceptor stacked polycyclic aromatic hydrocarbons as photocatalytic sites in carbon dots. *Nanoscale* **17**, 22834–22848 (2025).
2. Jia, L. *et al.* Predicting redox potentials by graph-based machine learning methods. *J. Comput. Chem.* **45**, 2383–2396 (2024).
3. RDKit: Open-source cheminformatics. https://www.rdkit.org/.
4. 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).
5. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
6. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
7. Frisch, M. J. *et al.* Gaussian 16, Revision C.01, Gaussian, Inc., Wallingford CT, 2016.
8. Lin, Y.-S., Li, G.-D., Mao, S.-P. & Chai, J.-D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **9**, 263–272 (2013).
9. Grimme, S. A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules. *J. Chem. Phys.* **138**, 244104 (2013).
10. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).

11. Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput. Mol. Sci.* **12**, e1606 (2022).

12. Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **17**, 616–641 (1996).

13. Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **17**, 520–552 (1996).

14. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).

15. Chirlian, L. E. & Francl, M. M. Atomic charges derived from electrostatic potentials: A detailed study. *J. Comput. Chem.* **8**, 894–905 (1987).

16. Pracht, P., Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169–7192 (2020).

17. Hourahine, B. *et al.* DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **152**, 124101 (2020).

18. Gaus, M., Goez, A. & Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **9**, 338–354 (2013).

19. Kubillus, M., Kubař, T., Gaus, M., Řezáč, J. & Elstner, M. Parameterization of the DFTB3 Method for Br, Ca, Cl, F, I, K, and Na in Organic and Biological Systems. *J. Chem. Theory Comput.* **11**, 332–342 (2015).

20. Plasser, F. *et al.* Statistical analysis of electronic excitation processes: Spatial location, compactness, charge transfer, and electron-hole correlation. *J. Comput. Chem.* **36**, 1609–1620 (2015).

21. Bäppler, S. A., Plasser, F., Wormit, M. & Dreuw, A. Exciton analysis of many-body wave functions: Bridging the gap between the quasiparticle and molecular orbital pictures. *Phys. Rev. A* **90**, 052521 (2014).

22. Plasser, F., Wormit, M. & Dreuw, A. New tools for the systematic analysis and visualization of electronic excitations. I. Formalism. *J. Chem. Phys.* **141**, 024106 (2014).

23. Plasser, F. TheoDORE: A toolbox for a detailed and automated analysis of electronic excited state computations. *J. Chem. Phys.* **152**, 084108 (2020).