Electronic Supplementary Material (ESI) for Nanoscale.

Electronic Supplementary Material: Molecular generation targeting desired electronic properties via deep generative models

Qi Yuan,[†] Alejandro Santana-Bonilla,[†] Martijn A. Zwijnenburg,[‡] and Kim E. Jelfs[*,†]

†*Department of Chemistry, Molecular Sciences Research Hub, White City Campus, Imperial College London, Wood Lane, London, W12 0BZ, United Kingdom*

‡*Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom*

E-mail: k.jelfs@imperial.ac.uk

Phone: +44 (0)207 594 3438

Contents

1 Influence of theory level on the RNN prediction

In this section, we compare the results obtained by computing the electronic ground state and excited state properties employing the B3LYP [1–3] and ωB97X-D3 exchange-correlation functionals.[4] One of the crucial points in obtaining a successful transfer model is the quality and availability of examples from which the RNN can learn. Therefore, we wanted to quantify the impact on the quality of predictions made by the RNN when the different sets of data were used. We first performed a statistical analysis of the HOMO-LUMO gap energies and the dipole moment values displayed in Fig. S1(a,b). The results show a significant deviation from a linear relation, which can be ascribed to the different physical situations described by the exchange-correlation functionals in which the correct long-range decay is considered only in the ωB97X-D3 exchange-correlation.[4] Similarly, one can observe a broad distribution of values for both the HOMO-LUMO gap and the dipole moment values, displaying the variety of oligomers considered in the transfer database from the Computational Materials Repository.[5] It is noteworthy to mention that different DFT approaches will result in different predictions of the RNN, making it important to select an adequate level of theory for the deep generative models according to the investigated system.

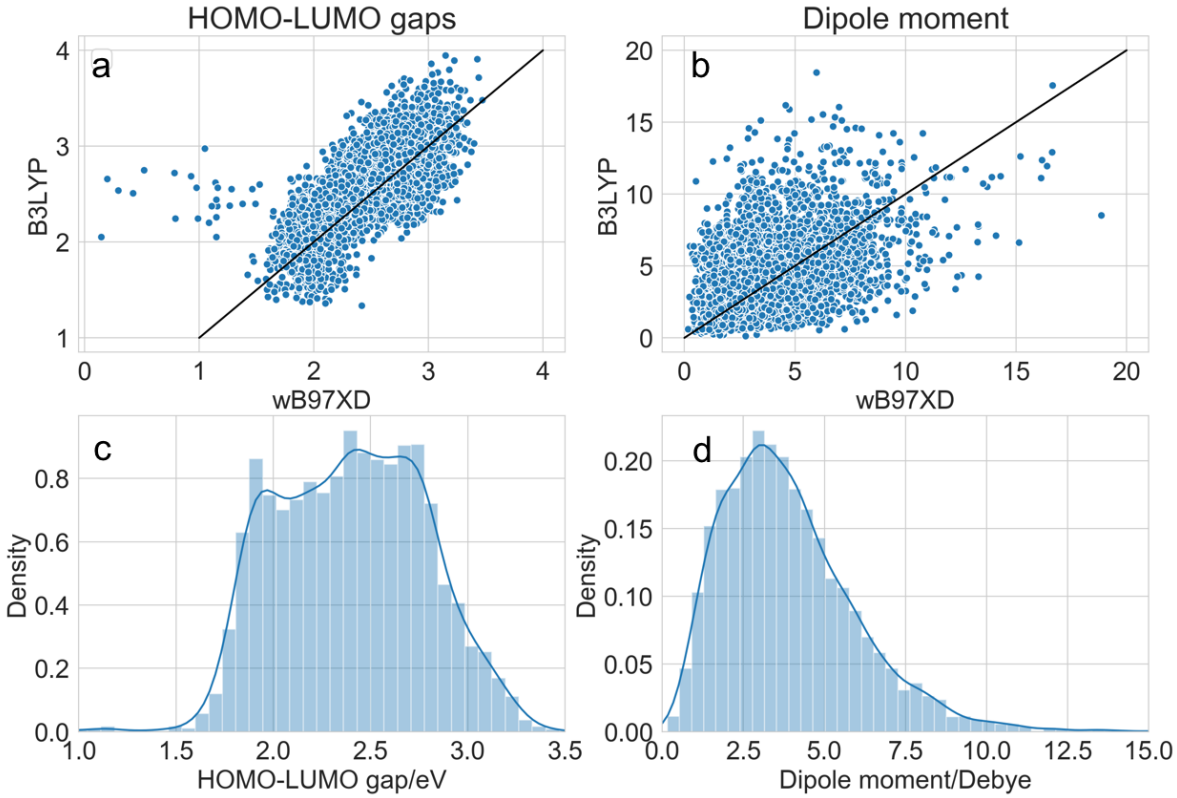

Figure S1: Comparison of the HOMO-LUMO gaps (a) and dipole moments (b) of oligomers in the Transfer Database computed using the ωB97X-D3 and B3LYP functionals. Distribution of HOMO-LUMO gaps (c) and dipole moments (d) of oligomers in the Transfer Database calculated using the ωB97X-D3 functional.

2 Performance of the General RNN and different TL models

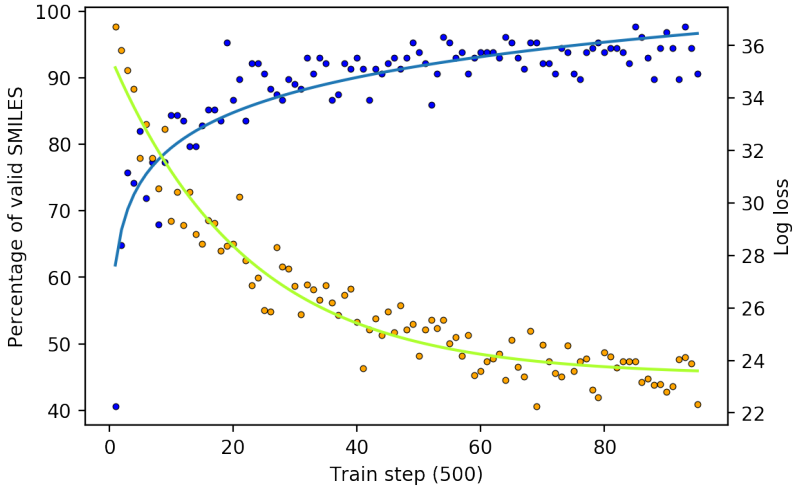

Figure S2: Percentage of valid SMILES generated using the General RNN (blue) and the log loss of the General RNN (green/orange) along with the training steps.

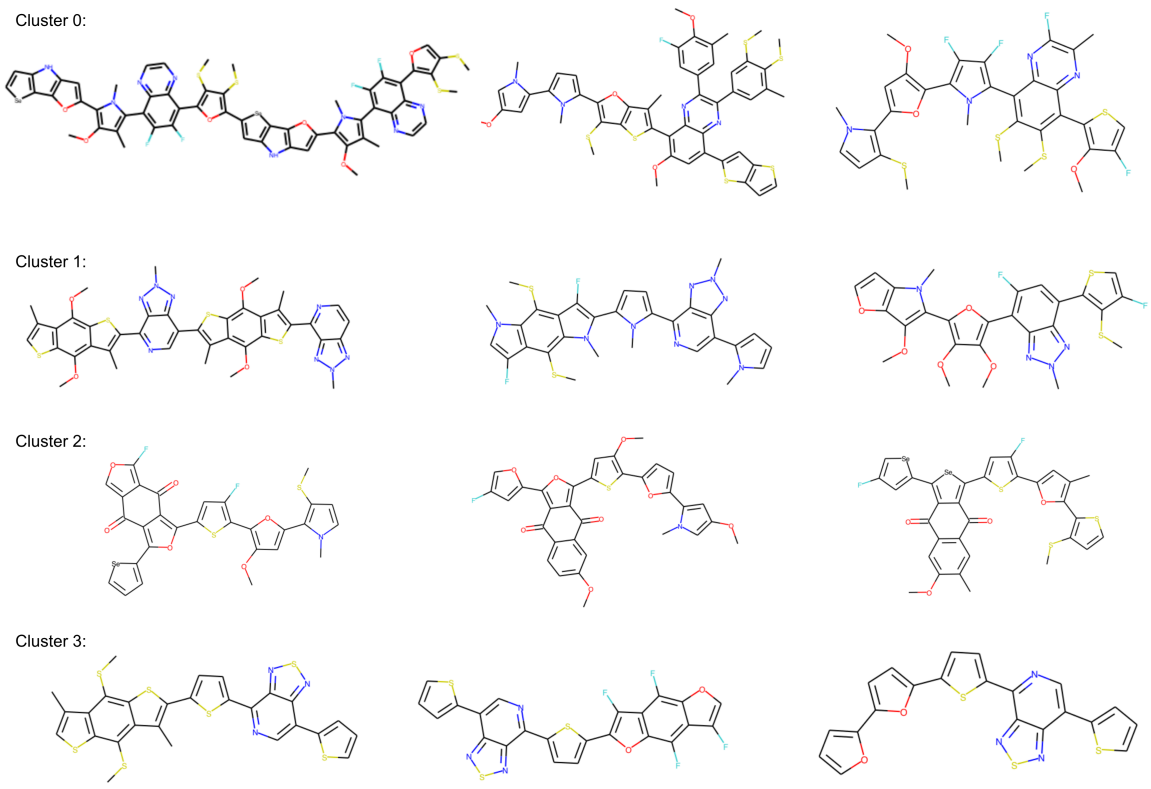

Figure S3: Molecular structures of representative molecules from the four clusters detected by Louvain method in the Transfer Database.

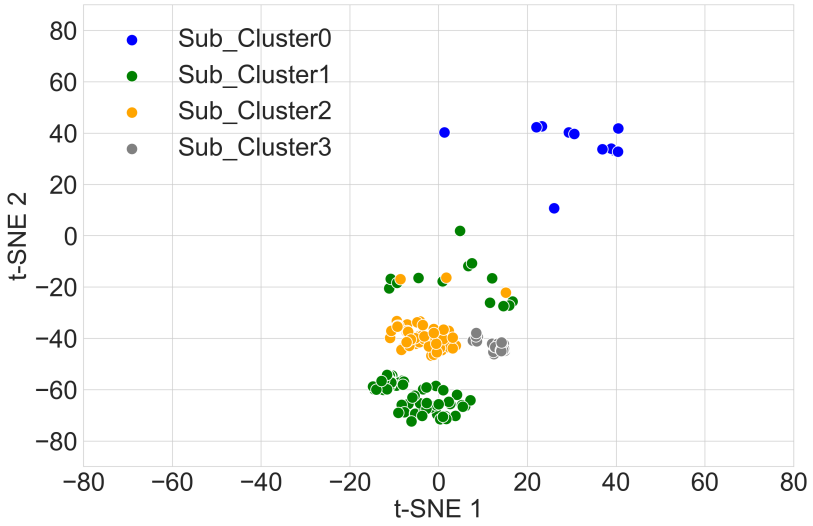

Figure S4: T-SNE projection of the molecular fingerprints of the 269 'promising' oligomers in the Transfer Database. Molecules are coloured by the sub clusters they belong to.

Table S1: Summary of performance of the six transfer learning models, with the performance metrics measured on the 15,360 SMILES strings sampled by each model.

Model	**Valid**	**Unique**	**Novel**	**Promising**
1	9019	6703	4346	456
2	11141	4756	3400	1060
3	5749	4322	4047	115
4	7959	4481	4252	2547
5	6592	4727	4635	1730
6	6927	4241	4157	2056

3 Learning structural and electronic properties

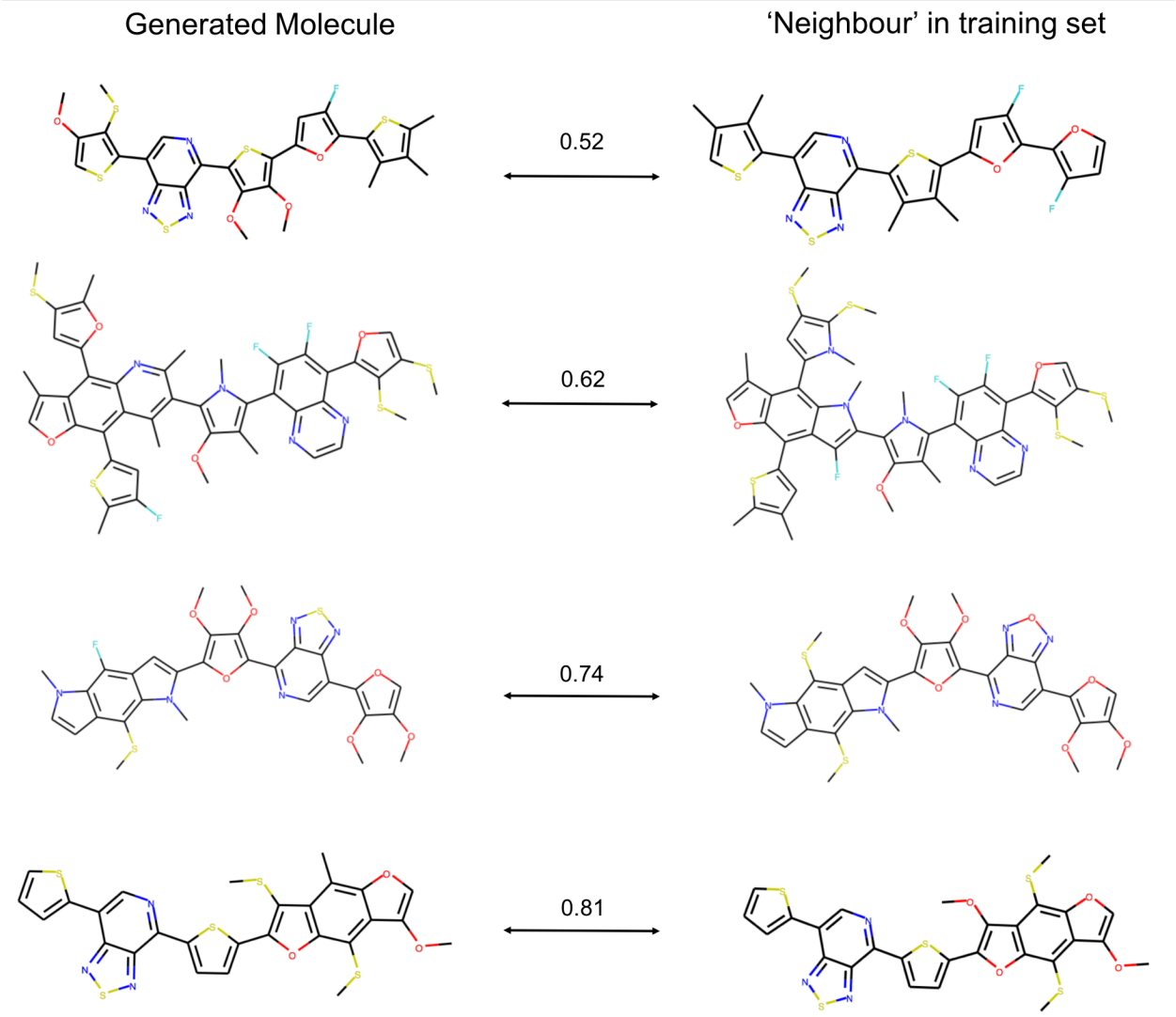

Figure S5: Examples of generated oligomers with low and high Tanimoto similarity to their neighbour in the Transfer Database

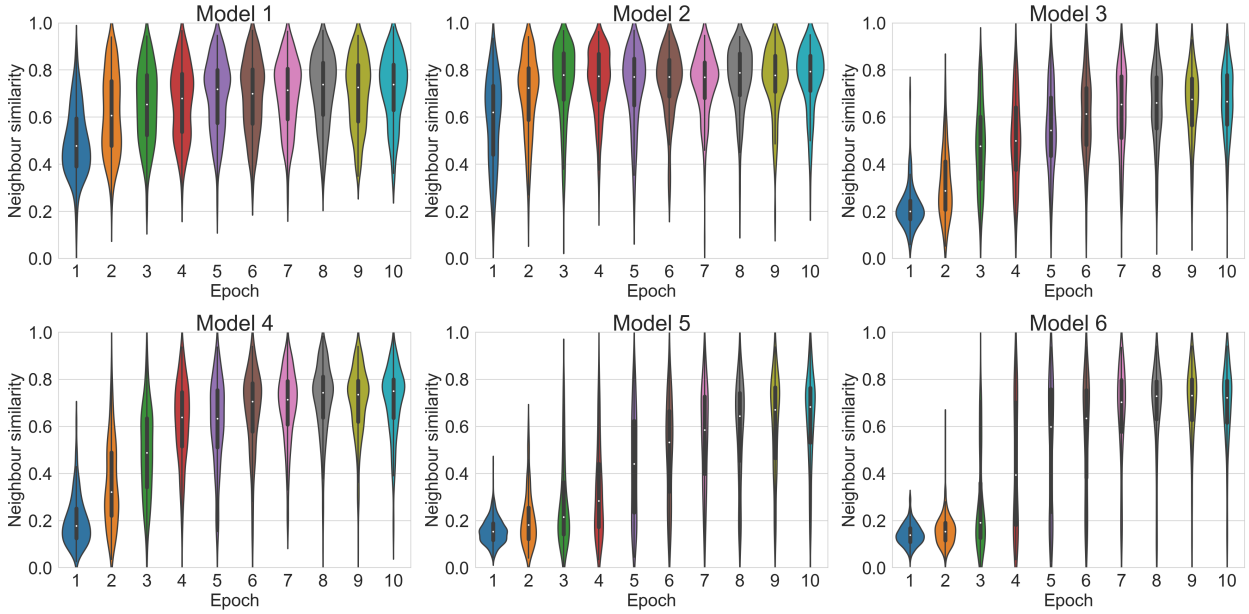

Figure S6: The distribution of the Tanimoto similarity of the Morgan fingerprints of the oligomers generated via transfer learning and their closest neighbour in the corresponding training set over 10 epochs.

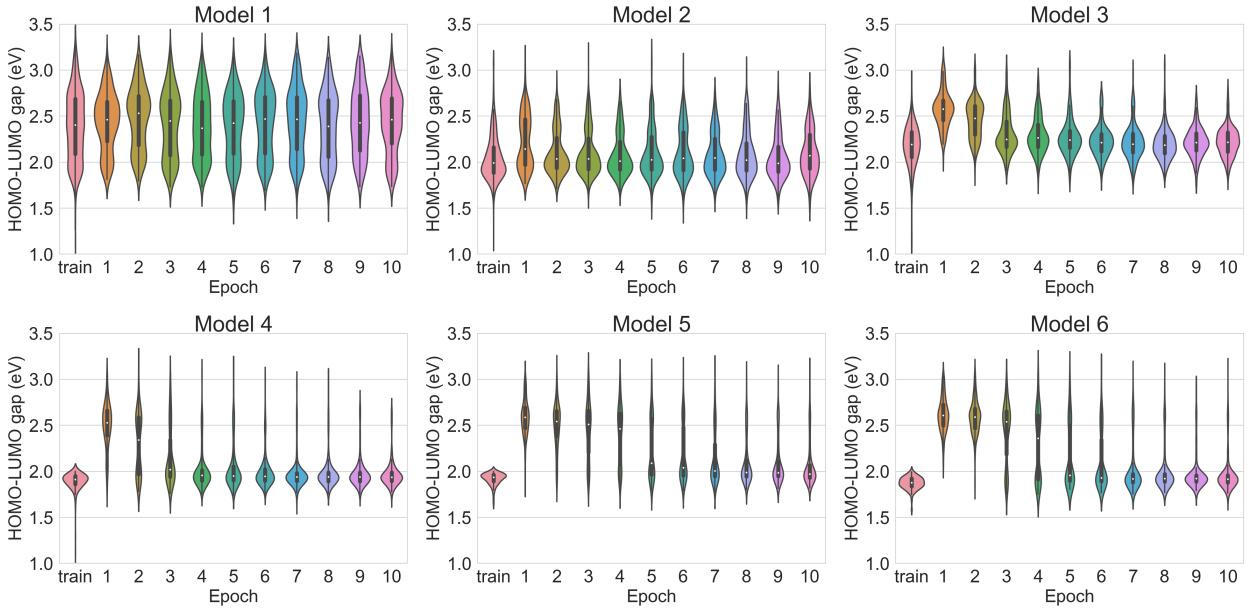

Figure S7: The distribution of the HOMO-LUMO gaps of oligomers generated via transfer learning over 10 epochs, together with the distribution of HOMO-LUMO gaps of the corresponding training set.

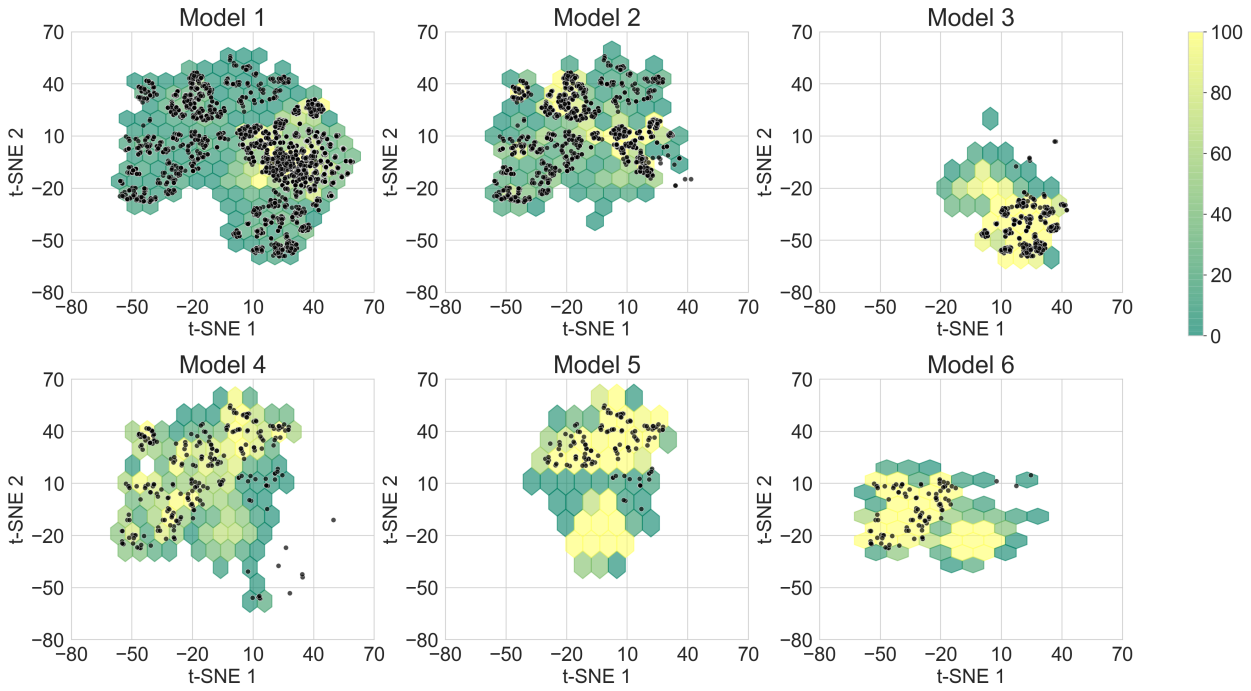

Figure S8: Hexagonal binning plot of t-SNE projection of the Morgan fingerprints of the oligomers generated from the TL models. Colours of the hexagons represent the number of generated oligomers in each hexagon according to the colour bar. The oligomers in the corresponding training sets are shown as scatterpoints in black in each sub plot. t-SNE 1 and t-SNE 2 corresponded to the first and second dimensions of the 2-D projection of the Morgan fingerprints.

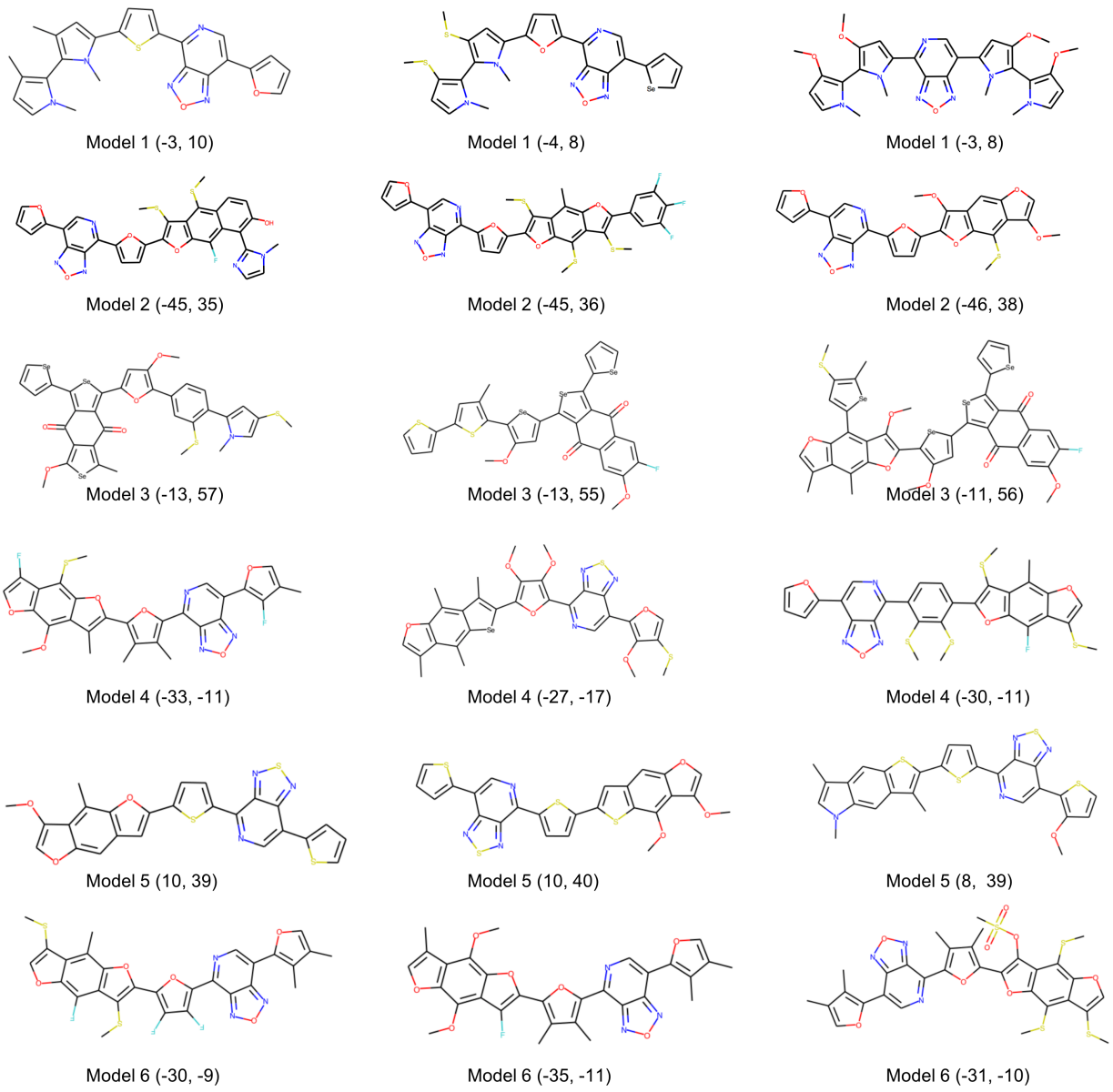

Figure S9: Molecules sampled from the six TL models, corresponding to different regions of the t-SNE projection of the generated 'promising' oligomers. Values in the parentheses represent the coordinates of the molecules in the t-SNE projection.

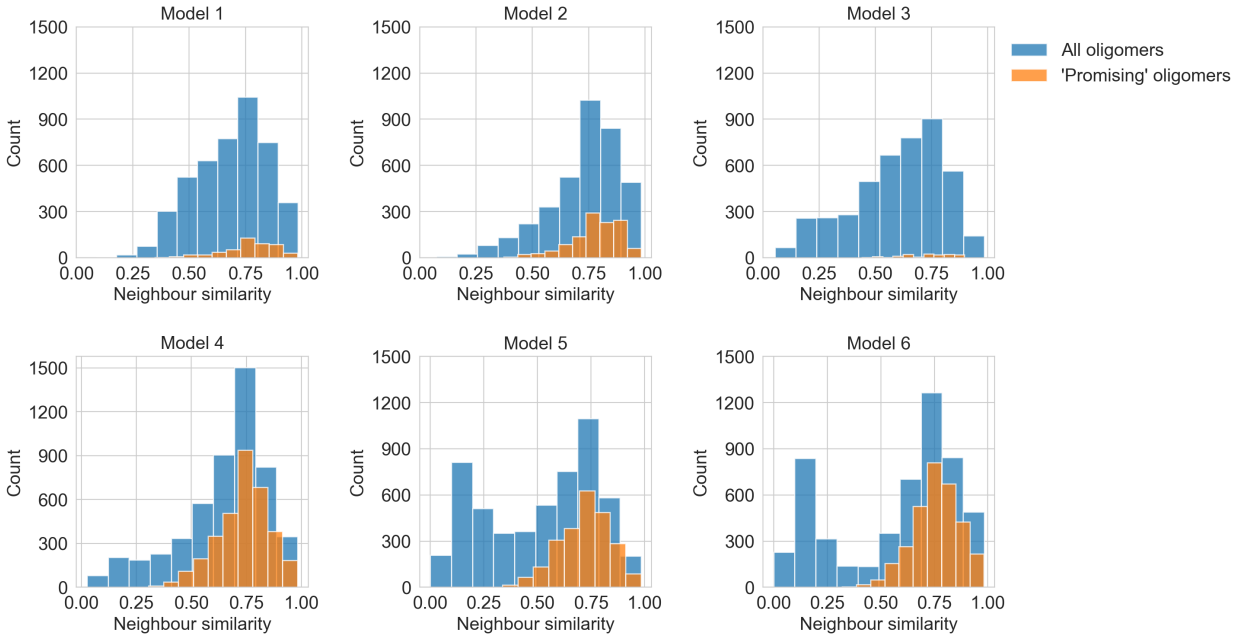

Figure S10: Histograms of the neighbour similarities of the molecules generated from the six TL models.

4 Promising donor-acceptor oligomers from the deep generative models

In this section, we report a reduced set of 22 'promising' oligomers predicted by our deep generative model. This group of selected oligomers has been optimized using the GFNx-TB2 code employing the same criteria as reported in the main text.[6] The optical and electronic properties (such as HOMO-LUMO gaps and dipole moments) were computed by employing the long-range corrected ωB97X-D3 functional[4] as implemented in ORCA[7] using the def2-TZVP basis set. Once the ground-state molecule was computed, we calculated the molecular electrostatic potential (MEP) to inspect the distribution of the charge. One can classify the oligomers into three different groups, whose main criterion is the extension of the homogeneous distribution of the electronic density. Based on this idea, one can see that molecules 7, 8, 11, 15, 16, 17, 18 and 19 in Fig. S12 can be classified into roughly homogeneous distributed electronic density over the moiety. Molecules 0, 1, 2, 3, 4, 5, 6, 10, 17, 20 and 21 exhibit a small degree of depletion and accumulation of charge at specific regions of the moiety. Finally, molecules 9 and 12 have a strong localisation of charge.

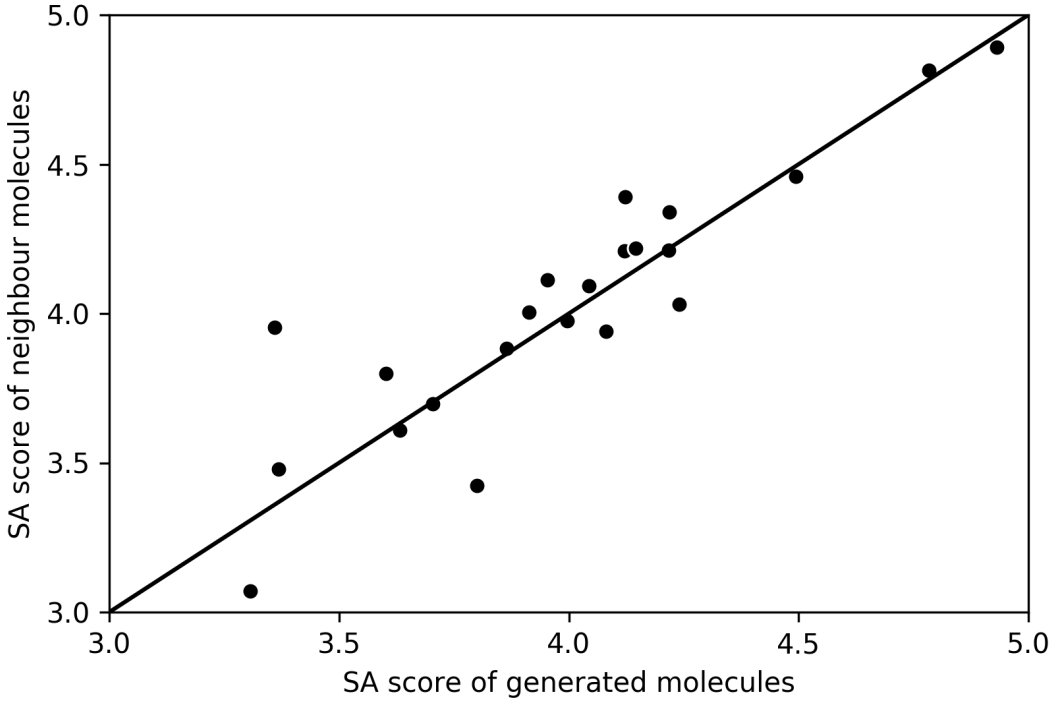

Figure S11: Comparison of SA scores (calculated according to Ertl and Schuffenhauer) of the 'promising' oligomers and their neighbours in the Transfer Database. An oligomer in the Transfer Database with the highest molecular similarity to a generated oligomer is referred as the 'neighbour' of the generated oligomer. The SA score measures the synthetic accessibility of molecules with a score between 1 (easy to make) and 10 (very difficult to make).

The optical properties have been computed at the same level of theory and by employing the simplified Tamm-Dancoff approach (sTDA) with an energy window of 6.5 eV for both singlet and triplet states.[8,9] The relevant computed quantities such as the optical HOMO-LUMO gaps and total dipole moments are collected and displayed in Table S2. As stated in the main text, the optical HOMO-LUMO gap and the total dipole moment can also be predicted by the deep generative model and directly compared with the corresponding expensive *ab initio* calculation results. We report a good accuracy of the value computed by the RNN model and the ones obtained by *ab initio* calculations for the electronic structure properties, while limitations in the correct prediction of dipole moments have been observed. Since the dipole moment is a quantity that relates to structural features (such as planarity and positions of functional groups in the space) and the electronic density distribution over the oligomer, it is challenging to describe well from a SMILES representation. In this particular subset of promising donor-acceptor oligomers, the produced candidates have achieved the imposed constraints in the HOMO-LUMO gap values. In the case of the dipole moments, the investigated oligomers are within the statistical error predicted from our analysis, introducing a possible source of error for a more efficient targeted search in the chemical space.

Finally, we have computed the absorption spectra for all 22 studied candidates and the results are presented in Fig. S13. As a general trend across all oligomers, the first and most intense bright state can be found in energy regions between 2 - 3 eV. In particular, molecule 10 exhibits the strongest intensity among all the studied moieties, which clearly indicates the influence of substitutions such as selenium in molecular scaffolds like BDT.[10] For excited states at higher energy (>3 eV), one can observe a general trend of moderate intensity absorption peaks up to energy regions of 6.5 eV. However, two exceptions to this behavior can be found among the chosen candidates. First, molecules 1, 12 and 19 have protruding intensities in an energy range between 3.5 - 5.0 eV, which are not seen in the other oligomers. On the contrary, in the case of molecule 15, at energies around 3 eV an

Table S2: Optoelectronic properties calculated with the (top row) sTDA-ωB97X-D3 functional and (bottom row) the RNN.

Molecule	HOMO-LUMO (eV)	Dipole moment (Debye)
mol_0	1.82	0.6
	1.75	2.3
mol_1	1.82	1.5
	1.98	2.3
mol_2	1.92	1.3
	1.89	1.9
mol_3	1.97	1.8
	1.95	1.8
mol_4	1.85	2.0
	1.96	3.4
mol_5	1.92	1.7
	1.94	1.3
mol_6	1.87	1.1
	1.95	2.4
mol_7	1.99	1.9
	1.92	2.6
mol_8	1.93	0.7
	1.89	1.9
mol_9	1.99	1.5
	1.97	3.6
mol_10	1.71	1.6
	1.70	2.6
mol_11	1.87	1.2
	1.88	2.4
mol_12	1.84	1.8
	1.88	1.7
mol_13	1.84	1.4
	1.81	2.3
mol_14	1.76	1.9
	1.99	2.4
mol_15	1.95	1.4
	1.90	1.6
mol_16	1.82	1.4
	1.97	2.9
mol_17	1.96	1.3
	1.92	2.2
mol_18	1.95	0.6
	1.89	2.4
mol_19	1.89	1.1
	1.88	2.5
mol_20	1.81	0.6
	1.77	1.7
mol_21	1.98	0.6
	1.99	2.2

optical band gap can be recognized.

The reported substitutions suggested by the RNN can be characterized in terms of chang-

ing capping groups such as hydrogen atoms by either halogenation or methyl groups, which are common strategies used traditionally to fine tune electronic structure features in donor-acceptor oligomers. However, recently experimental studies have focused upon controlling the molecular packing by introducing targeted substitutions in well established molecular scaffolds.[10] For example, in the serendipituous case of benzo[1,2-b:4,5-b]dithiophene (BDT), in which substitutions of sulfur by selenium atoms produced a positive impact on the measured charge-carrier mobility.[11] In our case, the deep generative model has also suggested this kind of substitution, as observed in molecules 8 and 16 in Fig. S12. This can be ascribed to an effective exploration of the chemical space encompassed in the transfer database in which the model has been trained.

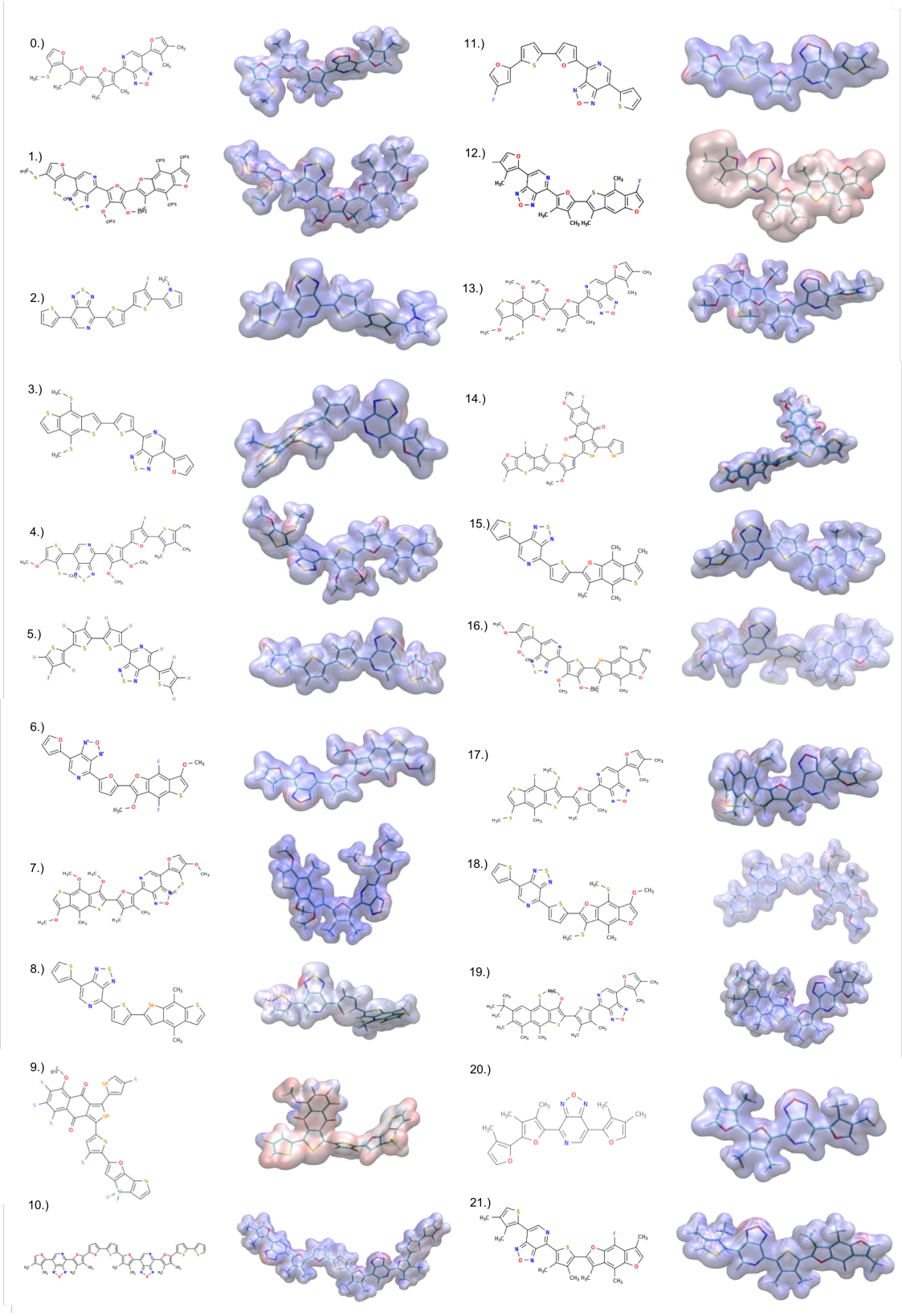

Figure S12: Schematic illustration of the promising candidates with 2-D depiction of the molecular structures, together with MEP in which red indicates electron localization whereas dark blue depicts electron depletion.

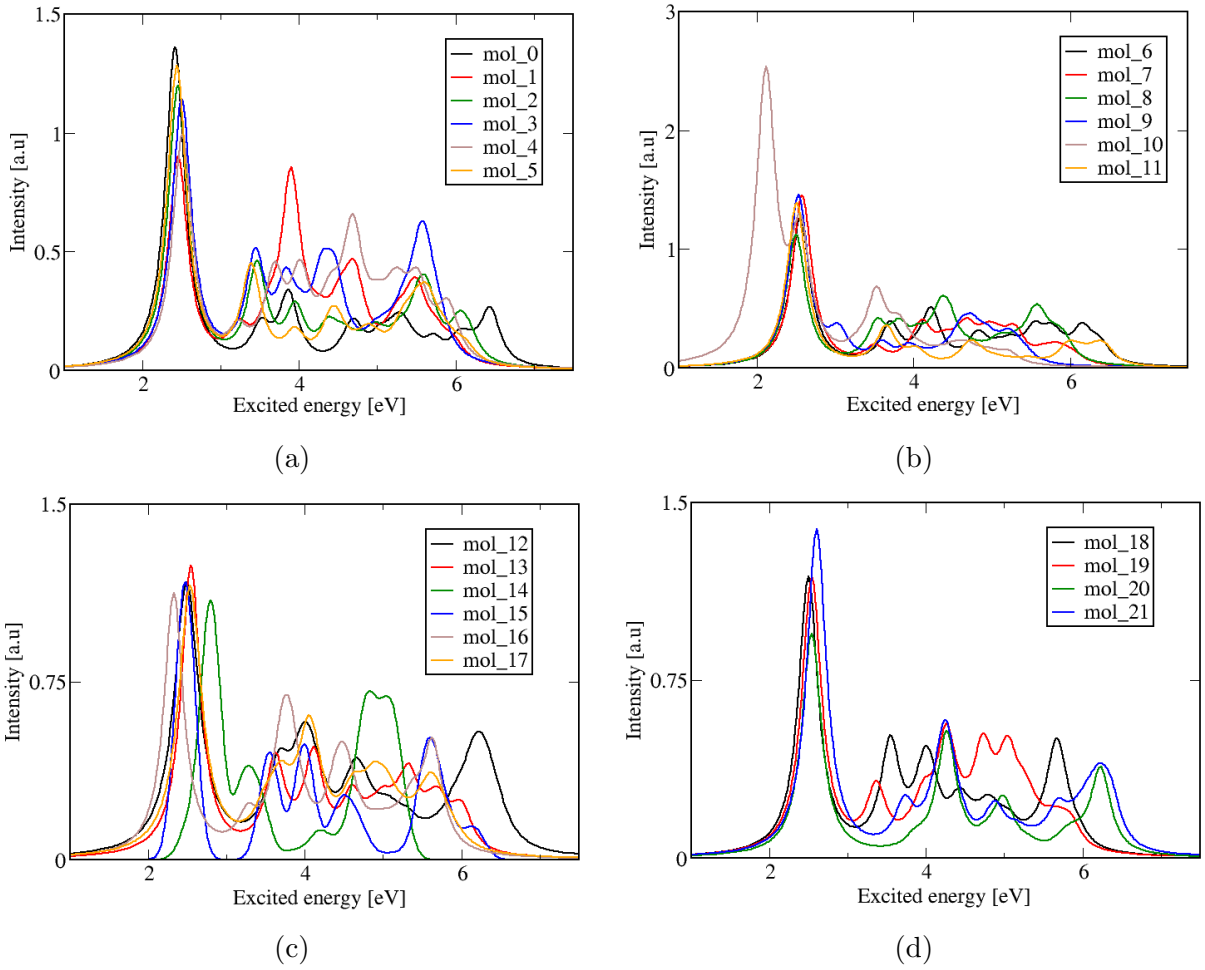

Figure S13: Computed UV-Vis spectra for the set of 22 'promising' oligomers employing the simplified Tamm-Dancoff approach (sTDA). The ωB97X-D3 exchange-correlation in combination with def2-TZVP basis set has been used to compute the displayed spectra.

5 Comparison of geometry optimized structures between GFNx-TB2 and ωB97X-D3

In this work, the 22 promising structures were optimized using ωB97X-D3 DFT level,[7] which yields very accurate equilibrium structures for a wide range of systems. ωB97X-D3/def2-TZVP was the reference structure level for the gas systems.[4,12,13] Subsequent GFNx-TB2[6] calculations were performed and both structures have been compared using the root-mean square displacement (RMSD) as indicative criterion of similarity. In general terms, there is a good agreement between the two methodologies, as displayed in Fig. S14, in which 15 oligomers out of the 'promising' 22 subset have an RMSD below 1.0 Å (*e.g.* molecule 11 in the right upper panel of Fig. S14). However, it is important to highlight that there are some cases in which GFNx-TB2[6] exhibits a bad performance, typically in candidates in which many heavy elements, such as sulphur, are present. This discrepancy can be ascribed to the fact that this methodology uses a minimal basis set which can not provide the sufficient flexibility to correctly describe these soft bonds.

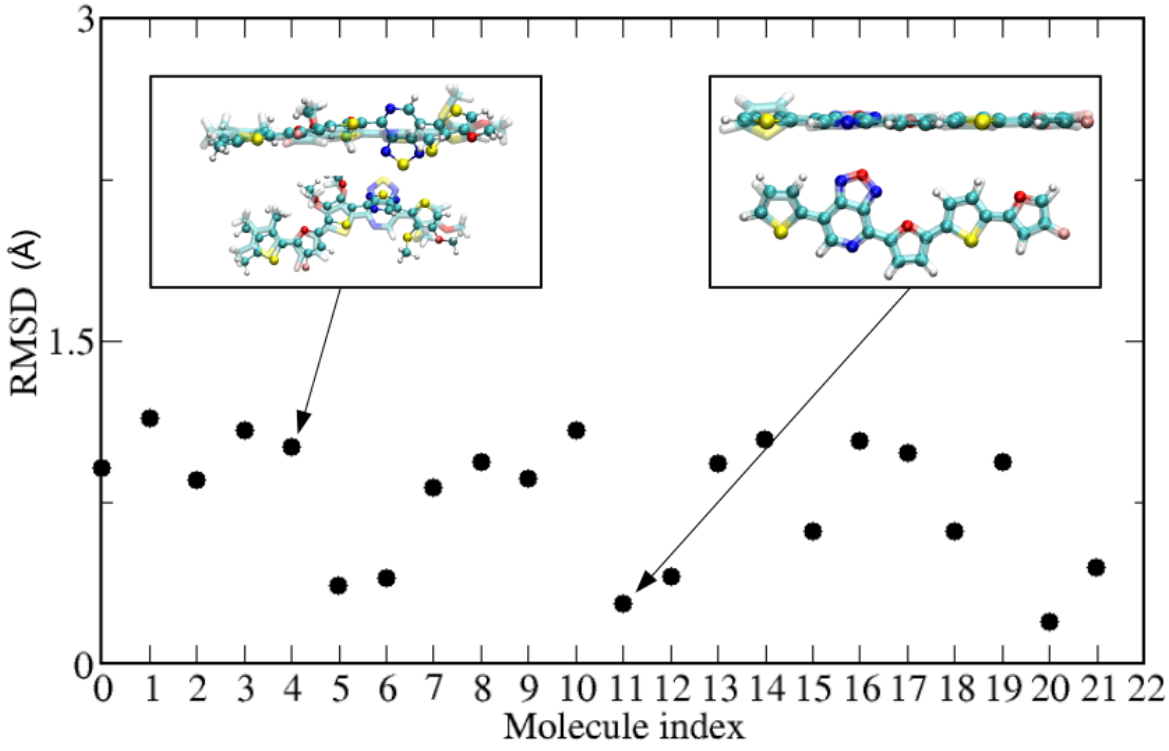

Figure S14: A comparison of structures optimized using GFNx-TB2 and ωB97X-D3/def2-TZVP.

References

(1) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(2) Raghavachari, K. Perspective on "Density functional thermochemistry. III. The role of exact exchange". *Theor. Chem. Acc.* **2000**, *103*, 361–363.

(3) Stephens, P. J.; Devlin, F.; Chabalowski, C.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(4) 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.* **2013**, *9*, 263–272.

(5) `https://cmr.fysik.dtu.dk/`.

(6) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(7) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1327.

(8) Grimme, S. A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules. *J. Chem. Phys.* **2013**, *138*, 244104.

(9) Bannwarth, C.; Grimme, S. A simplified time-dependent density functional theory approach for electronic ultraviolet and circular dichroism spectra of very large molecules. *Comput. Theor. Chem.* **2014**, *1040-1041*, 45 – 53, Excited states: From isolated molecules to complex environments.

(10) Yao, H.; Ye, L.; Zhang, H.; Li, S.; Zhang, S.; Hou, J. Molecular design of benzodithiophene-based organic photovoltaic materials. *Chem. Rev.* **2016**, *116*, 7397–7457.

(11) Takenaka, H.; Ogaki, T.; Wang, C.; Kawabata, K.; Takimiya, K. Selenium-Substituted β-Methylthiobenzo [1, 2-b: 4, 5-b] dithiophenes: Synthesis, Packing Structure, and Transport Properties. *Chem. Mater.* **2019**, *31*, 6696–6705.

(12) 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.* **2005**, *7*, 3297–3305.

(13) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.