Electronic Supporting Information for "High-Throughput Virtual Screening of Second-Order Nonlinear Optical Chromophores within the Donor-π-bridge-Acceptor Framework"

Chunyun Tu,[*,†] Weijiang Huang,[†] Sheng Liang,[‡] Kui Wang,[†] Qin Tian,[†] and Wei Yan[*,†]

[†]*School of Chemistry and Materials Engineering, Guiyang University, Guiyang, 550005, P. R. of China.*

[‡]*School of Mathematics and Information Science, Guiyang University, Guiyang, 550005, P. R. of China.*

E-mail: yidapa@sina.cn; lrasyw@163.com

Phone: +86 182 7500 1640

Generation of compound library

The compound library is constructed by combining donors, π-bridges and acceptors at preset connection sites (symbol * is used to denote the connection site) within the donor-π-acceptor frameworks. The structures of 30 donors, 21 π-bridges and 43 acceptors are shown in Figure **S1**, **S2** and **S3**, respectively.

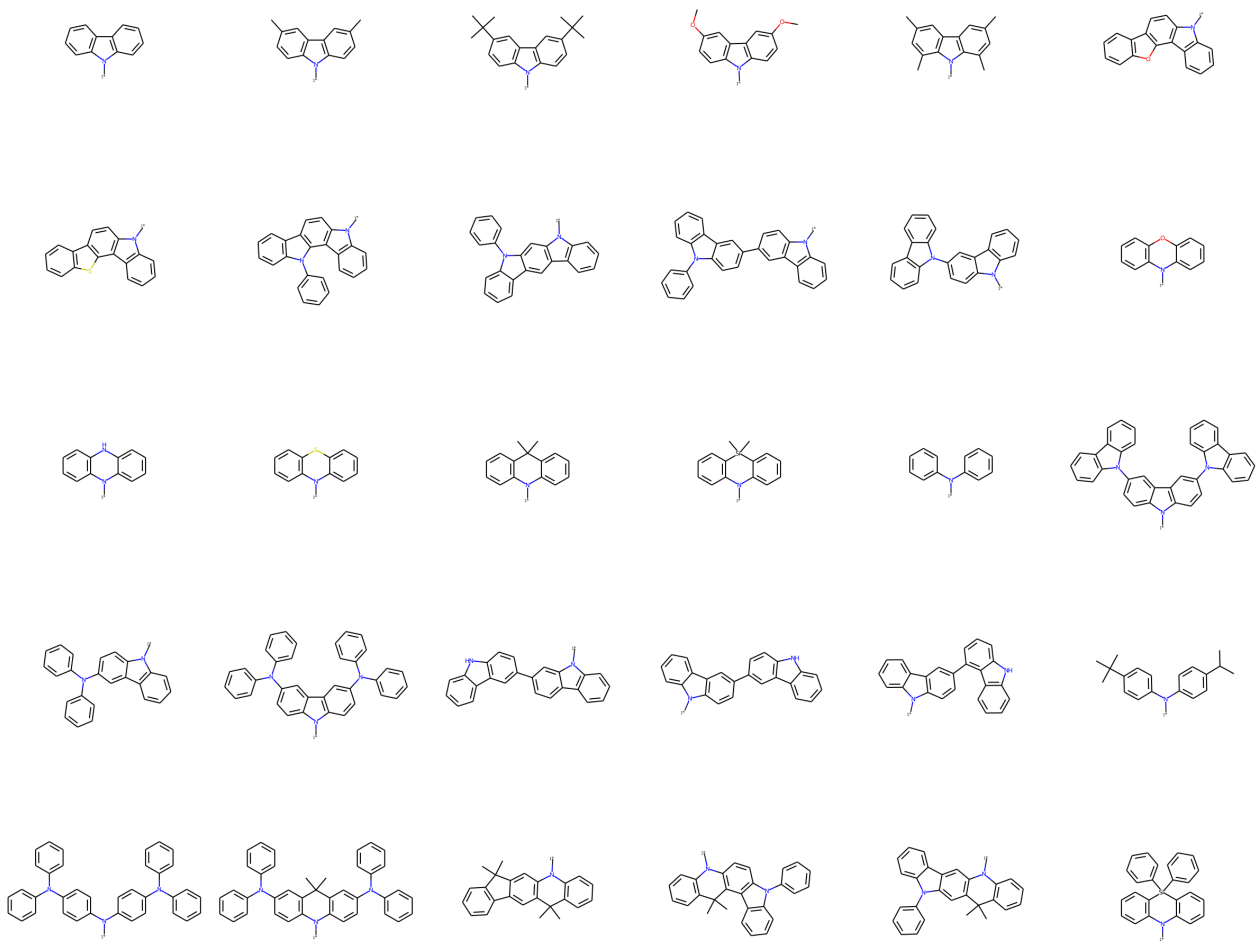

Figure S1: The donors (D) used as fragments for construction of donor-π-acceptor (D-π-A) molecules. (* is used to denote the connection site)

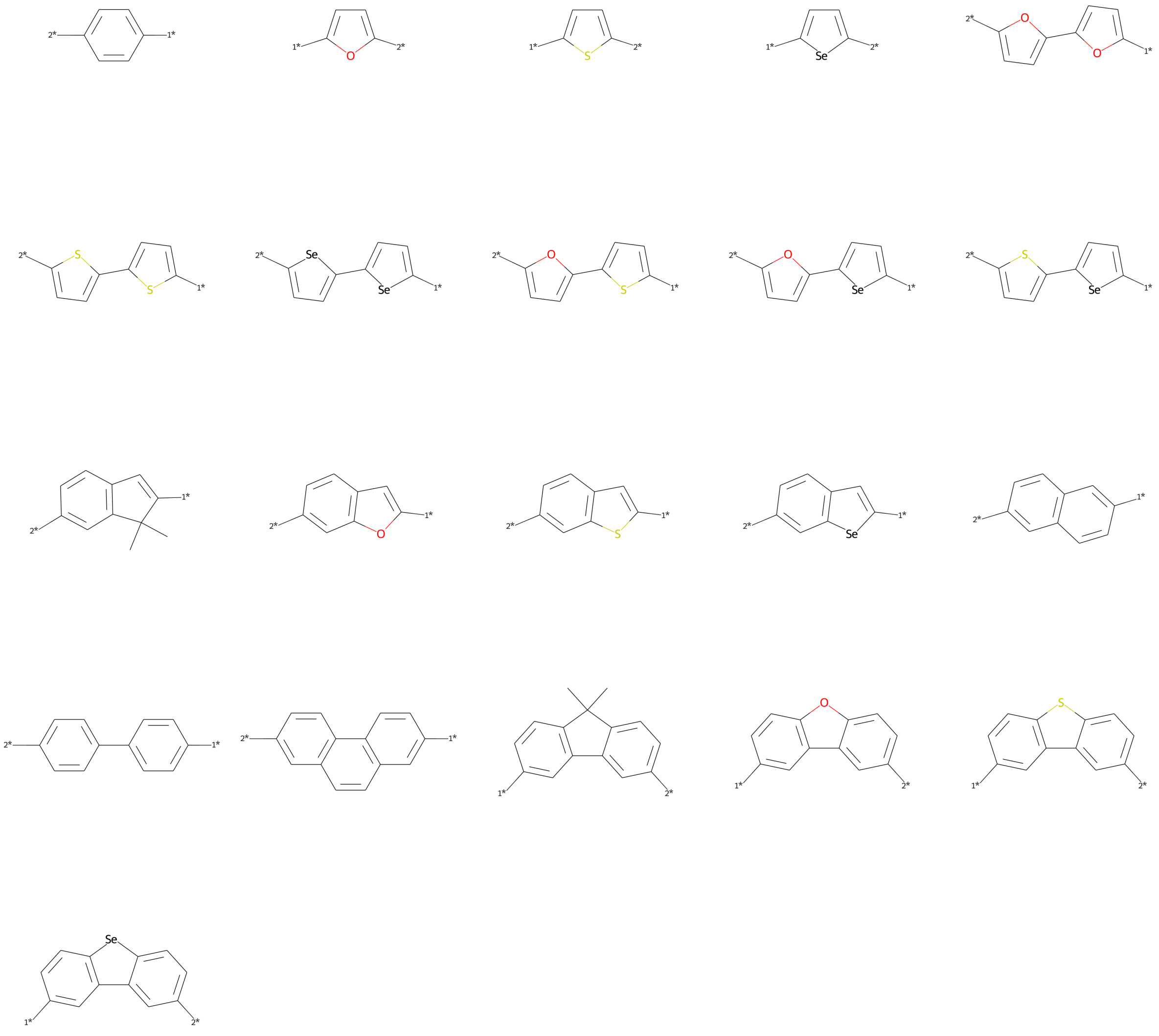

Figure S2: The π-bridges (π) used as fragments for construction of donor-π-acceptor (D-π-A) molecules. (1* and 2* are used to denote the connection sites)

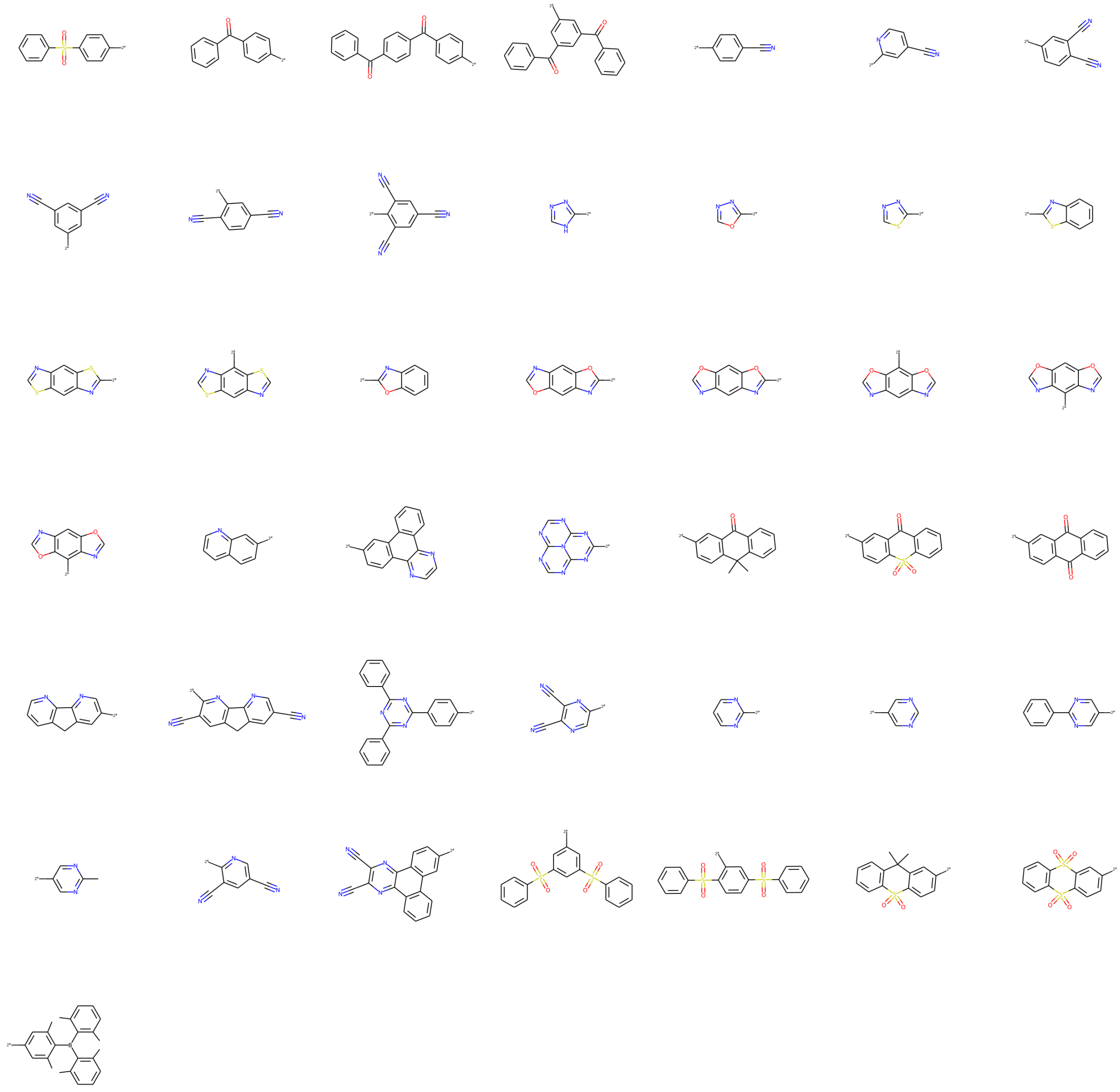

Figure S3: The acceptors (A) used as fragments for construction of donor-acceptor (DA) molecules. (* is used to denote the connection site)

Details for training and evaluation of machine learning model

The featurization of molecular structures for training molecules is carried out by utilizing the ECFP fingerprint (size = 2048) computing tool of the DeepChem package. By introducing the related tools of the open source machine learning Scikit-Learn package, a Random Forest Regressor (RandomForestRegressor, RF) from the ensemble module is adopted as the ML model, Grid Search Cross Validation (GridSearchCV) and relevant score function and method (cross_val_score and neg_mean_squared_error) as tools for model selection, simple imputer (SimpleImputer) as data imputer, and a min max scaler (MinMaxScaler) for data scaling. The following grid parameters have been used for the Grid Search Cross Validation step: ‘bootstrap’: [True, False], ‘n_estimators’: [3, 10, 30, 100], ‘criterion’: [“mse”, “mae”], ‘max_depth’: [2, 5, 10, 50], ‘max_features’: [“auto”, “sqrt”, “log2”]

The ECFP fingerprints as the X featurization vector, and the computed energy gaps (ΔE_{ST}) as the Y object vector. The X;Y is fed to the RF model, by applying the GridSearchCV with above grid parameters, the cross_val_score method and 5-fold cross validation, the best ML model (best_reg) is screened out from the grid search hyper-parameter space. The best_reg is retrained with the training data, and is further evaluated by a 5-fold cross validation using the same scoring method (neg_mean_squared_error). The newly learned ML model will be used for subsequent prediciting property of unseen molecules in the original compound library.

The training of the multi-layer perceptron (MLP) regressor is very similar as above, with the following grid parameters for the Grid Search Cross Validation step: ’hidden_layer_sizes’: [(50,), (100,), (250,), (500,), (50, 50), (100, 50), (100, 100), (250,50), (250,100), (250,250), (500, 50), (500, 100), (500, 250), (500, 500)], ’activation’: [’identity’, ’logistic’, ’tanh’, ’relu’], ’learning_rate_init’: [0.01, 0.005, 0.001].

Obviously, the mean (about 10^3) and standard deviation of some of the models are small enough (as compared with the large spanning range of the property, $10^0 \sim 10^5$), hence the ML model could be safely used to predict the first hyperpolarizability of unseen molecules with considerable confidence.

Distribution of Murcko decomposition skeletons

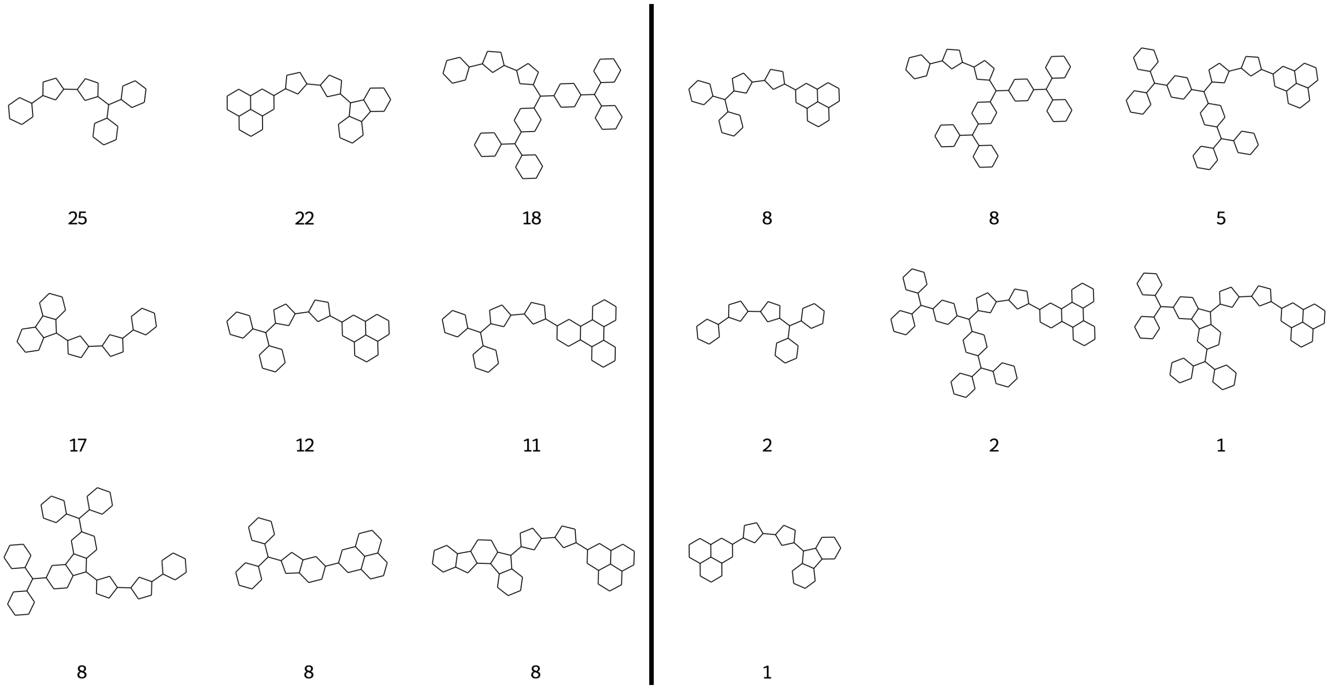

Figure S4: The structures and frequencies of occurrence for generic cores for top 1% (left side) and 1‰ (right side).

Solvent effect on NLO response

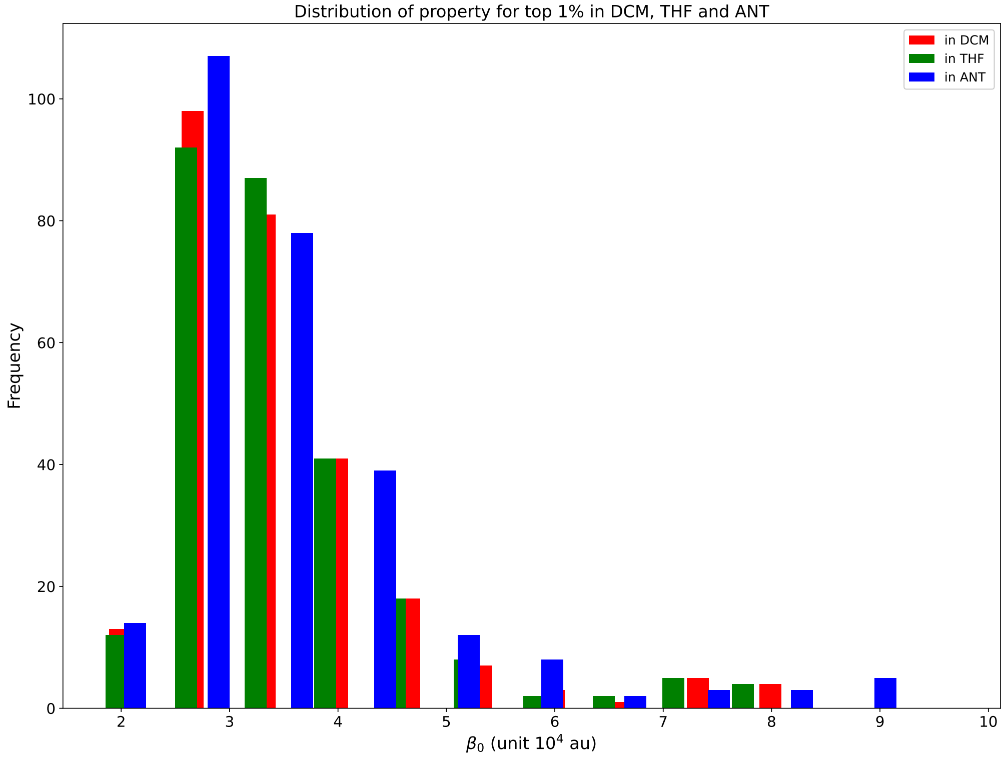

Figure S5: The distribution of calculated hyperpolarizability for top 1% molecules in solvents with different polarity. (DCM=dichloromethane, THF=tetrahydrofuran and ANT=acetonitrile)

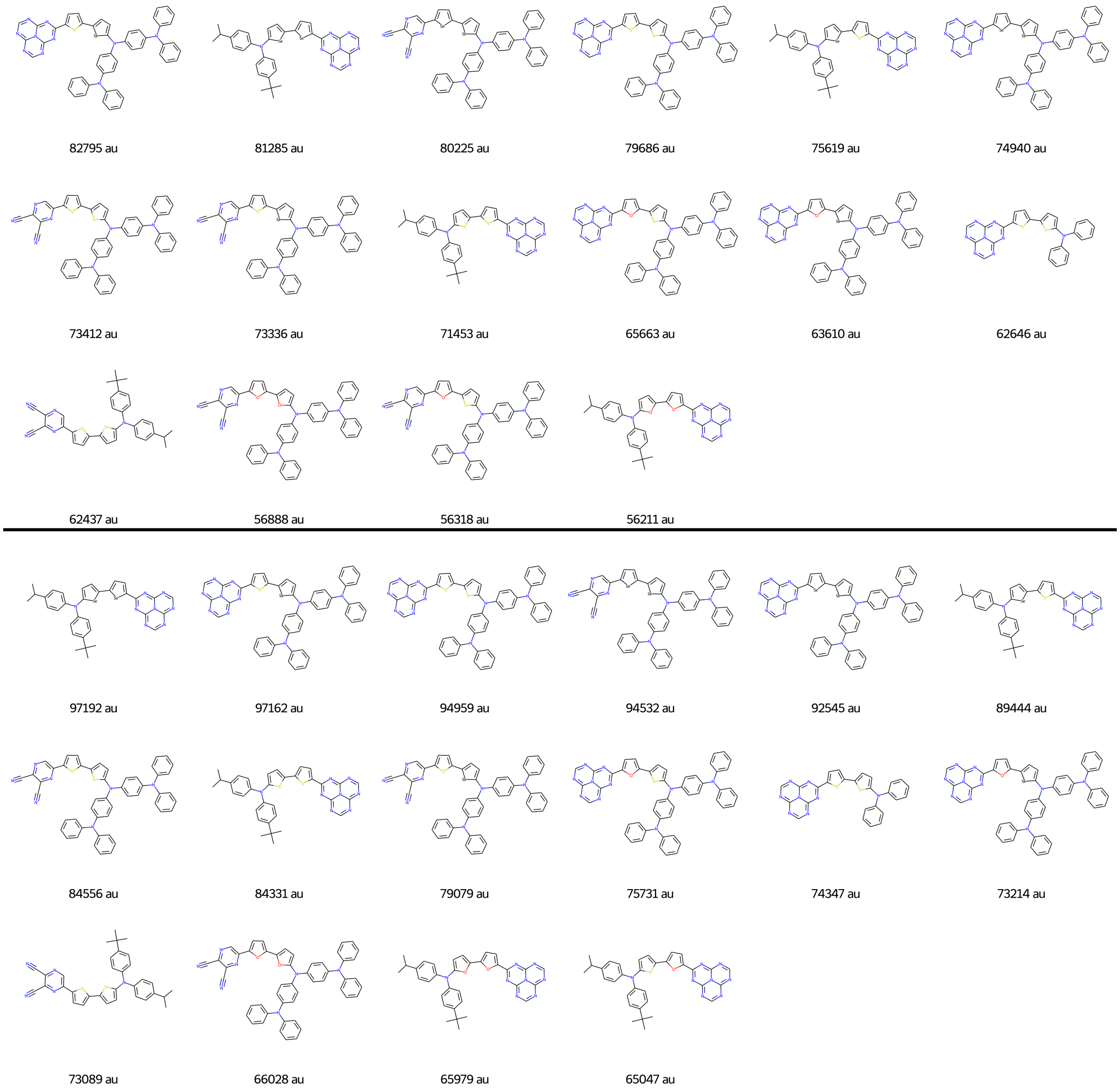

Figure S6: The structures of 16 optimal molecules with largest calculated static first hyperpolarizability in THF (on top) and ANT (at bottom). (THF=tetrahydrogenfuran and ANT=actonitrile)

Molecules with additional π spacers

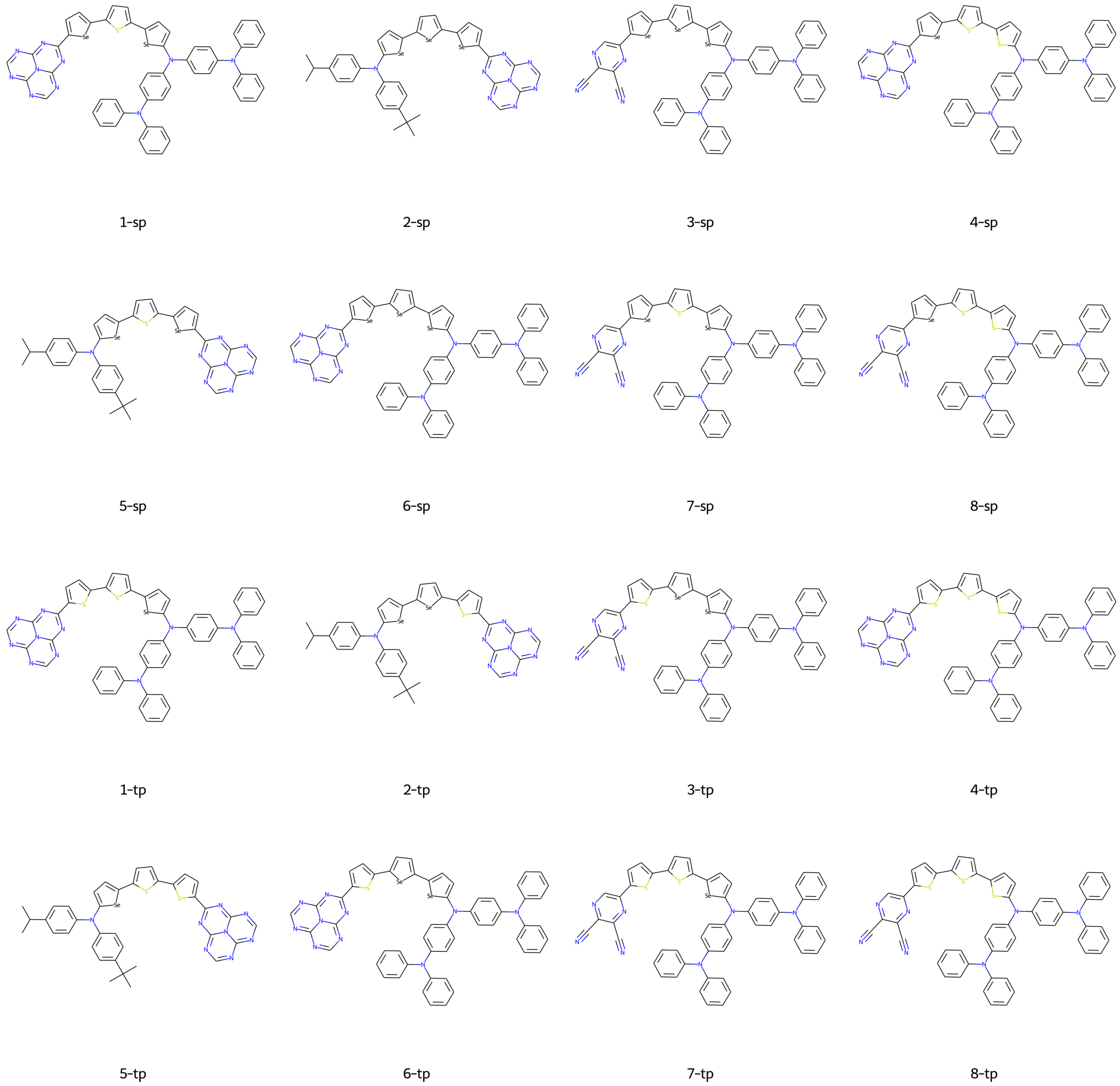

Figure S7: The structures of 8 top molecules with additional π spacers.

Molecules with varied acceptors

Table S1: The first vertical excitation energy (E_{S1}), oscillator strength (f), dipole moment of ground-state (μ_g), dipole moment difference between S_0 and S_1 ($\Delta\mu$), two-level theory estimated static first hyperpolarizability ($\beta_0(est)$), computed hyperpolarizability (β_0), and transferred charge (q_{CT}) and distance (d_{CT}) of charge-transfer between S_0 and S_1 of molecules with varied acceptors.

compounds	E_{S1} (eV)	E_{S1} (nm)	f	μ_g (Debye)	$\Delta\mu$ (Debye)	$\beta_0(est)$ (au)	β_0 (au)	q_{CT} (e)	d_{CT} (Å)
DEA-PHV-QXL-TCF	2.16	575	2.497	31.96	24.94	4.90E+04	2.01E+05	0.832	6.31
DEA-PHV-QXL-TCP	1.79	693	2.222	27.25	30.23	9.29E+04	4.45E+05	0.889	7.11
DPA-L2SeSe-TCF	1.89	657	2.151	32.03	11.48	2.90E+04	1.78E+05	0.667	3.90
DPA-L2SeSe-TCP	1.57	788	2.248	33.56	8.07	3.72E+04	2.82E+05	0.581	3.12
DPA-L2SeSe-TRZ	2.25	552	1.667	13.28	15.58	1.81E+04	0.92E+05	0.728	4.48
DPA-L2SeSe-TRZ-CN	2.11	586	1.759	19.32	16.31	2.42E+04	1.23E+05	0.730	4.64
DPA-L2SeSe-TRZ-$(CN)_2$	1.99	622	1.800	24.26	16.39	2.97E+04	1.53E+05	0.717	4.82
DPA-L2SeSe-TRZ-NO_2	2.08	595	1.744	20.70	16.40	2.52E+04	1.28E+05	0.726	4.70
DPA-L2SeSe-TRZ-$(NO_2)_2$	1.94	639	1.759	26.95	16.38	3.13E+04	1.62E+05	0.710	4.82
DPA-L2SeS-TCF	1.92	644	2.098	31.12	11.82	2.78E+04	1.67E+05	0.679	4.04
DPA-L2SeS-TCP	1.59	781	2.175	32.02	9.40	4.03E+04	2.98E+05	0.599	3.53
DPA-L2SeS-TRZ	2.30	539	1.665	12.84	16.25	1.76E+04	0.86E+05	0.749	4.54
DPA-L2SeS-TRZ-CN	2.16	574	1.746	18.74	17.33	2.38E+04	1.16E+05	0.757	4.77
DPA-L2SeS-TRZ-$(CN)_2$	2.03	611	1.774	23.52	17.62	2.96E+04	1.48E+05	0.750	4.96
DPA-L2SeS-TRZ-NO_2	2.13	583	1.726	20.09	17.51	2.48E+04	1.21E+05	0.752	4.85
DPA-L2SeS-TRZ-$(NO_2)_2$	1.97	629	1.725	26.12	17.72	3.17E+04	1.58E+05	0.744	4.96
DPA-L2SSe-TCF	1.88	658	2.124	30.63	13.33	3.38E+04	1.91E+05	0.700	4.25
DPA-L2SSe-TCP	1.56	794	2.234	32.08	10.34	4.83E+04	3.37E+05	0.617	3.65
DPA-L2SSe-TRZ	2.23	555	1.616	12.39	17.36	2.01E+04	0.95E+05	0.763	4.75
DPA-L2SSe-TRZ-CN	2.10	591	1.710	18.75	17.88	2.62E+04	1.29E+05	0.766	4.92
DPA-L2SSe-TRZ-$(CN)_2$	1.97	629	1.751	23.46	18.35	3.33E+04	1.66E+05	0.762	4.98
DPA-L2SSe-TRZ-NO_2	2.07	599	1.700	20.14	17.88	2.72E+04	1.35E+05	0.761	4.94
DPA-L2SSe-TRZ-$(NO_2)_2$	1.91	648	1.712	26.12	18.37	3.58E+04	1.78E+05	0.744	5.14
DPA-L2SS-TCF	1.92	645	2.074	29.70	13.76	3.20E+04	1.77E+05	0.714	4.37
DPA-L2SS-TCP	1.58	786	2.163	30.55	11.67	5.08E+04	3.47E+05	0.639	4.08
DPA-L2SS-TRZ	2.29	542	1.609	11.90	18.05	1.92E+04	0.87E+05	0.782	4.79
DPA-L2SS-TRZ-CN	2.14	578	1.688	18.09	19.03	2.60E+04	1.21E+05	0.793	5.04
DPA-L2SS-TRZ-$(CN)_2$	2.00	619	1.715	22.67	19.71	3.35E+04	1.59E+05	0.787	5.20
DPA-L2SS-TRZ-NO_2	2.11	588	1.671	19.50	19.19	2.71E+04	1.28E+05	0.788	5.12
DPA-L2SS-TRZ-$(NO_2)_2$	1.94	639	1.668	25.29	19.85	3.60E+04	1.73E+05	0.777	5.32
DPA-L3SeSeSe-TRZ-$(CN)_2$	1.81	685	2.080	23.52	23.57	6.56E+04	3.17E+05	0.828	5.87
DPA-L3SeSeS-TRZ-$(CN)_2$	1.85	670	2.031	22.57	24.67	6.28E+04	2.90E+05	0.848	6.03
DPA-L3SeSSe-TRZ-$(CN)_2$	1.85	671	2.039	22.85	24.46	6.25E+04	2.92E+05	0.843	6.06
DPA-L3SeSS-TRZ-$(CN)_2$	1.89	657	1.989	21.97	25.60	5.98E+04	2.68E+05	0.862	6.16
DPA-L3SSeSe-TRZ-$(CN)_2$	1.82	682	2.016	22.80	24.97	6.62E+04	3.11E+05	0.845	6.20
DPA-L3SSeS-TRZ-$(CN)_2$	1.86	667	1.962	21.94	26.03	6.30E+04	2.83E+05	0.863	6.31
DPA-L3SSSe-TRZ-$(CN)_2$	1.86	667	1.964	21.88	26.07	6.31E+04	2.82E+05	0.865	6.29
DPA-L3SSS-TRZ-$(CN)_2$	1.90	654	1.913	21.23	27.13	6.00E+04	2.59E+05	0.883	6.40

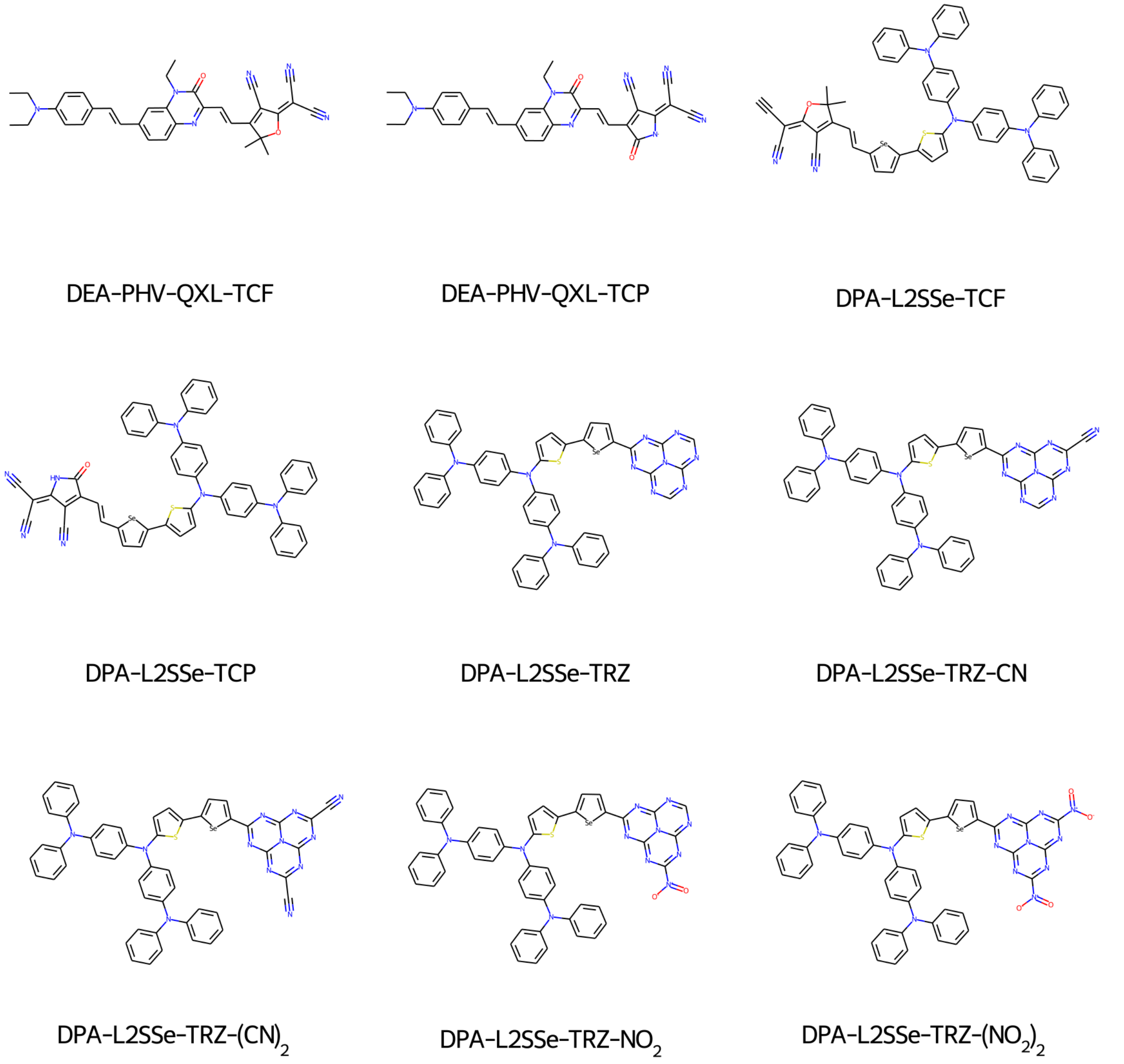

Figure S8: The structures and shorthand names of representative molecules with varied acceptors.