Supporting Information

Transfer Learning Accelerated Discovery of Conjugated Oligomers for Advanced Organic Photovoltaics

*Siyan Deng, Jing Xiang Ng, and Shuzhou Li**

Siyan Deng - School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore.

Jing Xiang Ng - School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore.

Corresponding Author

Shuzhou Li - School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore; E-mail: lisz@ntu.edu.sg

1. Dataset selection

PubChemQC is one of the largest quantum chemistry databases available. From this extensive dataset, we selected a subset known as PubChemQC-100K, comprising approximately 100,000 molecules (106,429) with more than six double bonds and a HOMO-LUMO gap below 6 eV. This specific selection was made to increase the similarity between the base dataset and the target dataset, which is crucial for developing robust pre-trained models. The size of the PubChemQC-100K dataset, with approximately 100,000 data points, is ideal for training robust pre-trained models. This volume of data is large enough to capture a wide variety of molecular features and interactions, enabling the deep learning model to learn complex patterns effectively. A well-trained model on such a dataset can generalize better to new, unseen data, which is crucial for the high-throughput screening process. Despite its large size, the PubChemQC-100 dataset is not excessively large to cause impractically long training times. The dataset size strikes a balance, being large enough to ensure robustness and diversity in the training process while remaining manageable in terms of computational resources and time required for training. This efficiency is critical for iterative model development and fine-tuning, allowing for faster experimentation and optimization cycles. The data within PubChemQC-100 is computed using the B3LYP/6-31G* level of theory, known for its high computational accuracy. This method strikes an excellent balance between computational cost and accuracy, providing reliable quantum chemical properties. While other quantum chemistry databases exist, many use lower-accuracy methods like PM6. The high precision of B3LYP/6-31G* ensures that the dataset is of superior quality, making it particularly suitable for training models aimed at high-fidelity property prediction.

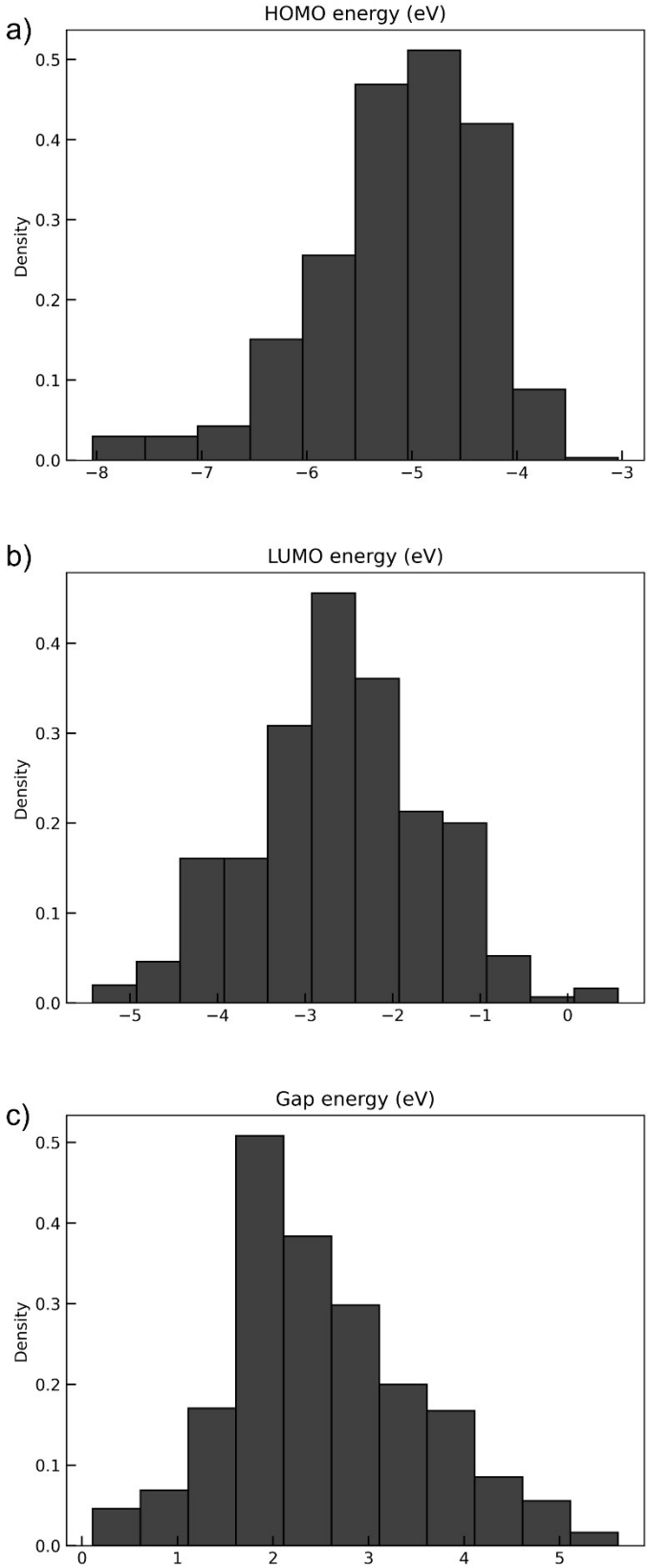


Figure S1. Distribution of HOMO, LUMO, and Gap for the CO-610 dataset.

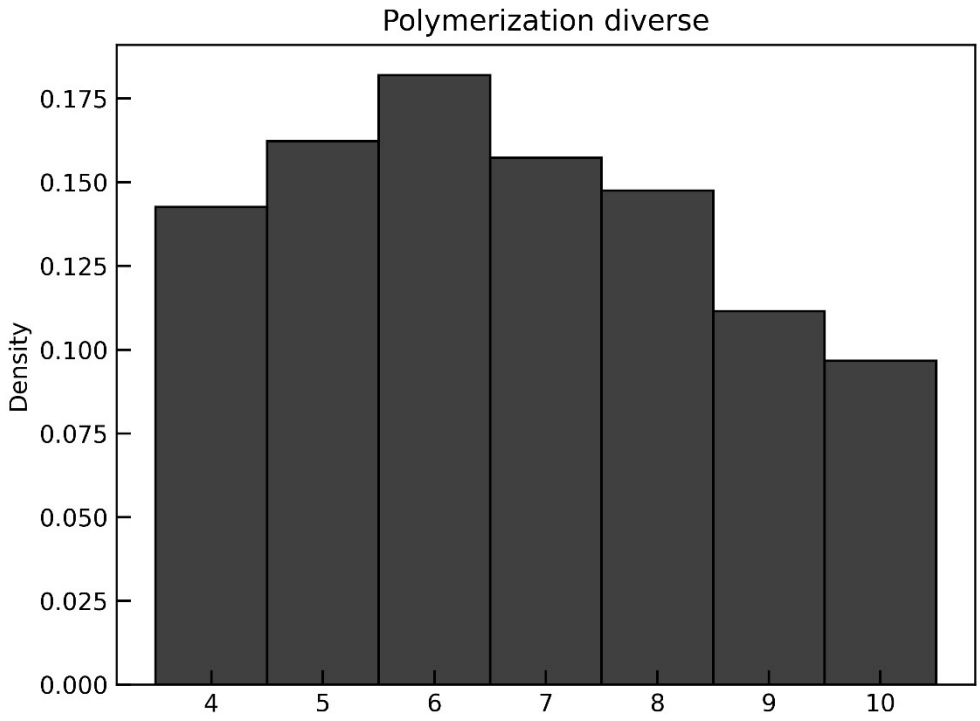
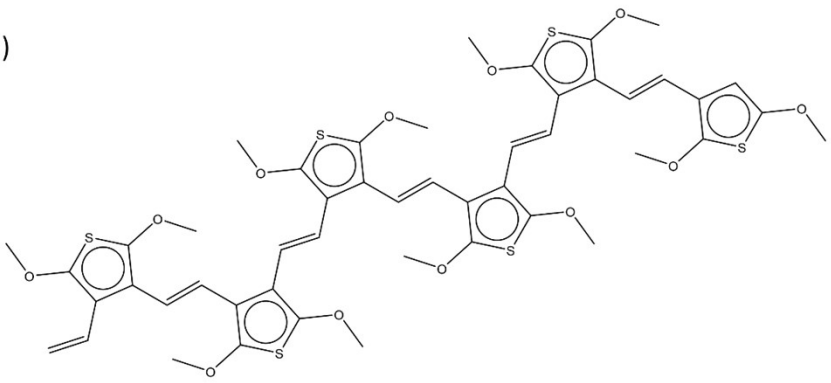


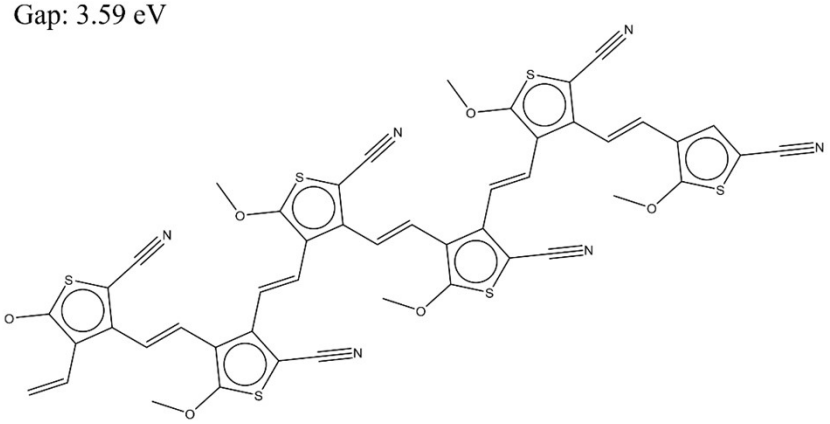
Figure S2. Distribution of polymerization degrees in the CO-610 dataset.

a)



HOMO: -4.56 eV
LUMO: -0.97 eV
Gap: 3.59 eV

b)



HOMO: -5.69 eV
LUMO: -2.49 eV
Gap: 3.20 eV

Figure S3. Oligomers with identical backbone structures exhibiting significantly different HOMO and LUMO levels.

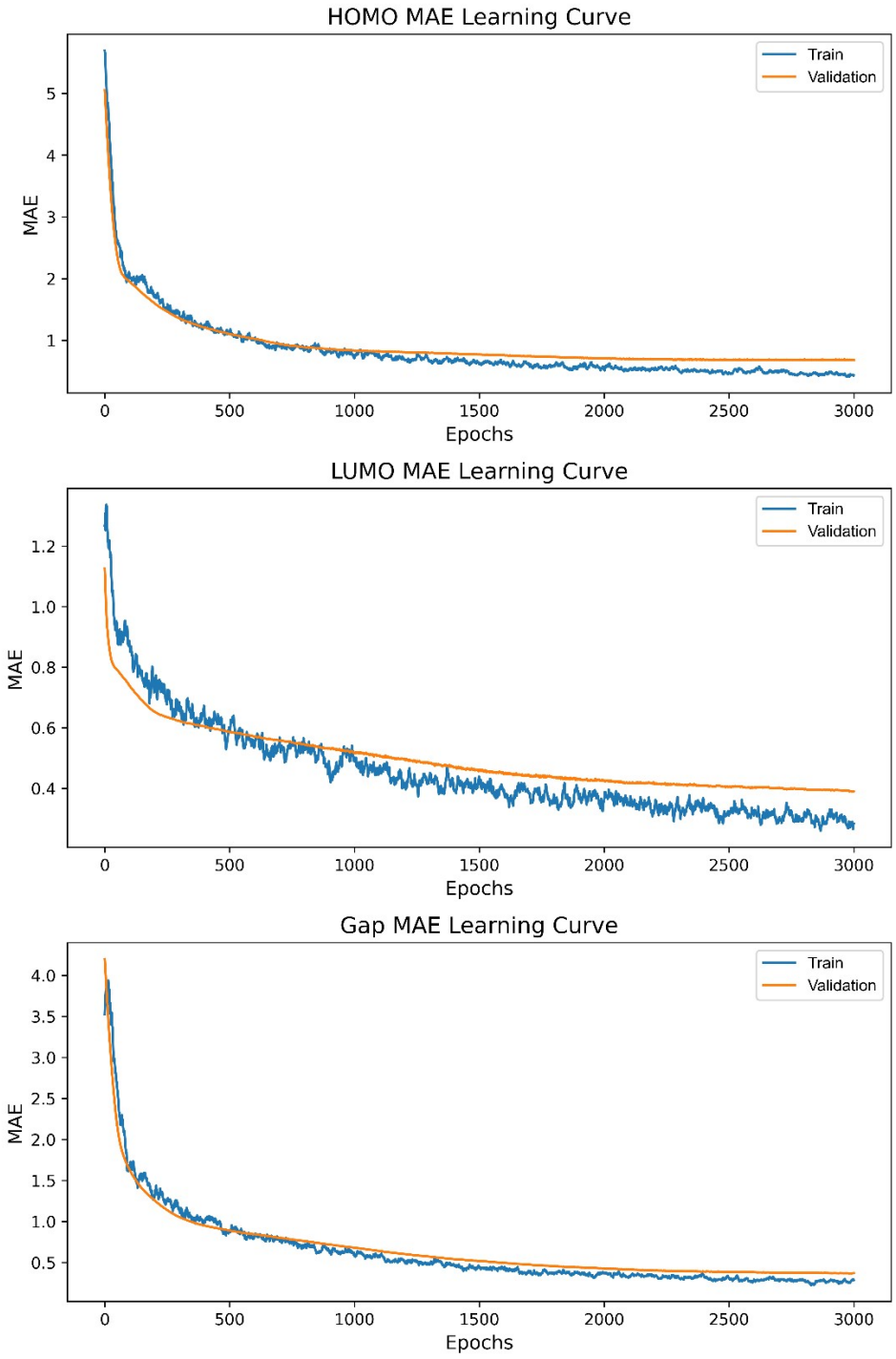


Figure S4. Learning curves for the training and validation sets for HOMO, LUMO, and Gap.

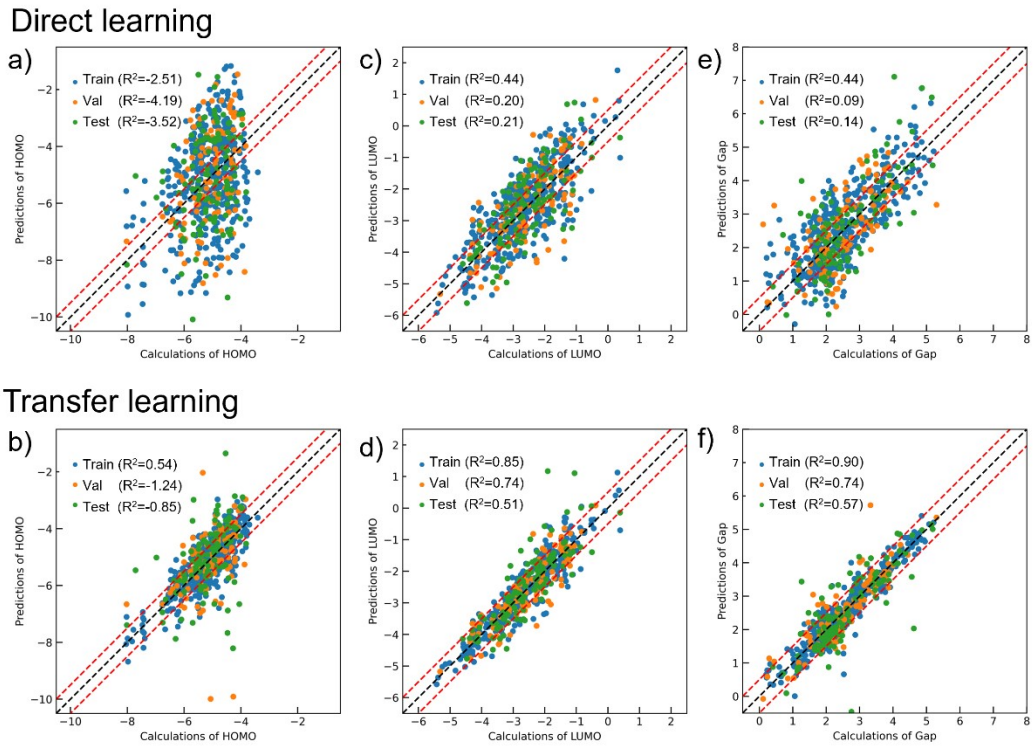


Figure S5. Comparison of predictions versus calculations for HOMO, LUMO, and Gap using direct learning and transfer learning approaches. (a), (c), and (e) represent predictions for HOMO, LUMO, and Gap, respectively, obtained through direct learning. (b), (d), and (f) show the corresponding predictions using transfer learning. Each plot includes data points for the training set (blue), validation set (orange), and test set (green). The black dashed line represents the ideal line, indicating perfect agreement between predictions and calculations. The red dashed lines represent deviations of ±0.5 eV from the ideal line.

Table S1. SMILES representations of candidates after DFT screening and their calculated HOMO, LUMO, and Gap.

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
104	5	C=Cc1ccnc1C=Cc1ccc(C=Cc2ccnc2C=Cc2ccc(C=Cc3ccnc3C=Cc3ccc(C=Cc4ccnc4C=Cc4ccc(C=Cc5ccnc5C=Cc5cccn5)n4)n3)n2)n1	-5.47	-3.72	1.74
117	7	C=CC=Cc1cnc(C=CC=Cc2cnc(C=CC=Cc3cnc(C=CC=Cc4cnc(C=CC=Cc5cnc(C=CC=Cc6cnc(C=CC=Cc7cncc7C)c6C)c5C)c4C)c3C)c2C)c1C	-5.29	-3.95	1.51
152	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9ccnc89)c8ccnc78)c7ccnc67)c6ccnc56)c5ccnc45)c4ccnc34)c3ccnc23)c2ccnc12	-5.51	-3.93	1.58
186	4	C=CC1=C2C(=O)NC(C=CC3=C4C(=O)NC(C=CC5=C6C(=O)NC(C=CC7=C8C(=O)NC=C8C(=O)O7)=C6C(=O)O5)=C4C(=O)O3)=C2C(=O)O1	-5.46	-4.19	1.27
187	10	c1cc2sc(-c3cc4sc(-c5cc6sc(-c7cc8sc(-c9cc%10sc(-c%11cc%12sc(-c%13cc%14sc(-c%15cc%16sc(-c%17cc%18sc(-c%19cc%20sccc%20n%19)cc%18n%17)cc%16n%15)cc%14n%13)cc%12n%11)cc%10n9)cc8n7)cc6n5)cc4n3)cc2n1	-5.71	-4.43	1.28
187	4	c1cc2sc(-c3cc4sc(-c5cc6sc(-c7cc8sccc8n7)cc6n5)cc4n3)cc2n1	-5.88	-4.03	1.85
188	4	C=Cc1cc2c(n1)C=C(C=Cc1cc3c(n1)C=C(C=Cc1cc4c(n1)C=C(C=Cc1cc5c(n1)C=CC5)C4)C3)C2	-5.10	-3.36	1.74
188	8	C=Cc1cc2c(n1)C=C(C=Cc1cc3c(n1)C=C(C=Cc1cc4c(n1)C=C(C=Cc1cc5c(n1)C=C(C=Cc1cc6c(n1)C=C(C=Cc1cc7c(n1)C=C(C=Cc1cc8c(n1)C=C(C=Cc1cc9c(n1)C=	-4.94	-3.60	1.34

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
		CC9)C8)C7)C6)C5)C4)C3)C2			
195	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc(OC)c7C#N)c(OC)c6C#N)c(OC)c5C#N)c(OC)c4C#N)c(OC)c3C#N)c(OC)c2C#N)c(OC)c1C#N	-4.93	-3.30	1.63
197	8	c1cc2nc3cc(-c4sccc5nc6cc(-c7sccc8nc9cc(-c%10sccc%11nc%12cc(-c%13sccc%14nc%15cc(-c%16sccc%17nc%18cc(-c%19sccc%20nc%21cc(-c%22sccc%23nc%24ccsc%24c%22-%23)sc%21c%19-%20)sc%18c%16-%17)sc%15c%13-%14)sc%12c%10-%11)sc9c7-8)sc6c4-5)sc3c-2cs1	-5.01	-3.10	1.91
199	4	C=CC1=C2C(=O)NC(C=CC3=C4C(=O)NC(C=CC5=C6C(=O)NC(C=CC7=C8C(=O)NC=C8C(=O)N7)=C6C(=O)N5)=C4C(=O)N3)=C2C(=O)N1	-5.11	-3.90	1.21
21	10	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12ccc(C(F)(F)F)cc%11%12)c%11ccc(C(F)(F)F)cc%10%11)c%10ccc(C(F)(F)F)cc9%10)c9ccc(C(F)(F)F)cc89)c8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1	-5.12	-3.26	1.86
21	7	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9ccc(C(F)(F)F)cc89)c8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1	-5.09	-3.17	1.92

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
21	8	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9scc%10ccc(C(F)(F)F)cc9%10)c9ccc(C(F)(F)F)cc89)c8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1	-5.07	-3.23	1.85
254	10	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9sc(C=Cc%10scc%11sc(=S)oc%10%11)c%10sc(=S)oc9%10)c9sc(=S)oc89)c8sc(=S)oc78)c7sc(=S)oc67)c6sc(=S)oc56)c5sc(=S)oc45)c4sc(=S)oc34)c3sc(=S)oc23)c2sc(=S)oc12	-5.30	-3.62	1.68
254	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9sc(=S)oc89)c8sc(=S)oc78)c7sc(=S)oc67)c6sc(=S)oc56)c5sc(=S)oc45)c4sc(=S)oc34)c3sc(=S)oc23)c2sc(=S)oc12	-5.30	-3.54	1.76
254	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10sc(=S)oc9%10)c9sc(=S)oc89)c8sc(=S)oc78)c7sc(=S)oc67)c6sc(=S)oc56)c5sc(=S)oc45)c4sc(=S)oc34)c3sc(=S)oc23)c2sc(=S)oc12	-5.30	-3.58	1.71
268	9	CCOC(=O)c1sc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11sc(C(=O)OCC)c(F)c%10%11)c%10sc(C(=O)OCC)c(F)c9%10)c9sc(C(=O)OCC)c(F)c89)c8sc(C(=O)OCC)c(F)c78)c7sc(C(=O)OCC)c(F)c67)c6sc(C(=O)OCC)c(F)c56)c5sc(C(=O)OCC)c(F)c45)c4sc(C(=O)OCC)c(F)c34)scc2c1F	-5.03	-3.10	1.93
272	10	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12cc(C#N)oc%11%12	-5.10	-3.97	1.13

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
		)c%11cc(C#N)oc%10%11)c%10cc(C#N)oc9%10)c9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1			
272	5	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6scc7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1	-5.33	-3.57	1.76
272	6	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1	-5.27	-3.71	1.57
272	7	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1	-5.24	-3.82	1.42
272	8	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9scc%10cc(C#N)oc9%10)c9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1	-5.16	-3.86	1.30
272	9	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11cc(C#N)oc%10%11)c%10cc(C#N)oc9%10)c9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1	-5.16	-3.95	1.20
295	5	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6scc7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1	-5.28	-3.50	1.78
295	6	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8cncnc78)c7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1	-5.21	-3.63	1.59

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
295	7	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9cncnc89)c8cncnc78)c7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1	-5.14	-3.75	1.39
328	5	C=CC#Cc1ccc(C=CC#Cc2ccc(C=CC#Cc3ccc(C=CC#Cc4ccc(C=CC#Cc5cccn5)n4)n3)n2)n1	-5.29	-3.95	1.51
351	10	C=Cc1cc(C=O)c(C=Cc2cc(C=O)c(C=Cc3cc(C=O)c(C=Cc4cc(C=O)c(C=Cc5cc(C=O)c(C=Cc6cc(C=O)c(C=Cc7cc(C=O)c(C=Cc8cc(C=O)c(C=Cc9cc(C=O)c(C=Cc%10cc(C=O)co%10)o9)o8)o7)o6)o5)o4)o3)o2)o1	-5.24	-3.27	1.97
373	8	C=Cc1cnc(C=Cc2cnc(C=Cc3cnc(C=Cc4cnc(C=Cc5cnc(C=Cc6cnc(C=Cc7cnc(C=Cc8cncc8C)c7C)c6C)c5C)c4C)c3C)c2C)c1C	-5.08	-3.73	1.35
382	10	O=Cc1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12ccc(C=O)cc%11%12)c%11ccc(C=O)cc%10%11)c%10ccc(C=O)cc9%10)c9ccc(C=O)cc89)c8ccc(C=O)cc78)c7ccc(C=O)cc67)c6ccc(C=O)cc56)c5ccc(C=O)cc45)c4ccc(C=O)cc34)scc2c1	-4.95	-3.37	1.58
382	9	O=Cc1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11ccc(C=O)cc%10%11)c%10ccc(C=O)cc9%10)c9ccc(C=O)cc89)c8ccc(C=O)cc78)c7ccc(C=O)cc67)c6ccc(C=O)cc56)c5ccc(C=O)cc45)c4ccc(C=O)cc34)scc2c1	-4.93	-3.34	1.58
394	9	CC(=O)c1cnc(-c2nc(-c3nc(-c4nc(-c5nc(-c6nc(-c7nc(-c8nc(-c9nccc9C(C)=O)cc8C(C)=O)cc7C(C)=O)cc6C(C)=O)cc5C(C)=O)cc4C(C)=O)cc3C(C)=O)cc2C(C)=O)c1	-5.29	-3.95	1.51

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
416	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc(O)c7C#N)c(O)c6C#N)c(O)c5C#N)c(O)c4C#N)c(O)c3C#N)c(O)c2C#N)c(O)c1C#N	-5.39	-3.42	1.98
416	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc(O)c8C#N)c(O)c7C#N)c(O)c6C#N)c(O)c5C#N)c(O)c4C#N)c(O)c3C#N)c(O)c2C#N)c(O)c1C#N	-5.37	-3.45	1.92
416	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc(O)c9C#N)c(O)c8C#N)c(O)c7C#N)c(O)c6C#N)c(O)c5C#N)c(O)c4C#N)c(O)c3C#N)c(O)c2C#N)c(O)c1C#N	-5.41	-3.45	1.96
421	4	C=Cc1ncc(C=Cc2ncc(C=Cc3ncc(C=Cc4nccc5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12	-5.47	-3.49	1.99
433	10	C=Cc1c(F)c(F)c(C=Cc2c(F)c(F)c(C=Cc3c(F)c(F)c(C=Cc4c(F)c(F)c(C=Cc5c(F)c(F)c(C=Cc6c(F)c(F)c(C=Cc7c(F)c(F)c(C=Cc8c(F)c(F)c(C=Cc9c(F)c(F)c(C=Cc%10c(F)c(F)cc%11nsnc%10%11)c%10nsnc9%10)c9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12	-5.33	-3.51	1.82
433	8	C=Cc1c(F)c(F)c(C=Cc2c(F)c(F)c(C=Cc3c(F)c(F)c(C=Cc4c(F)c(F)c(C=Cc5c(F)c(F)c(C=Cc6c(F)c(F)c(C=Cc7c(F)c(F)c(C=Cc8c(F)c(F)cc9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12	-5.35	-3.48	1.87
433	9	C=Cc1c(F)c(F)c(C=Cc2c(F)c(F)c(C=Cc3c(F)c(F)c(C=Cc4c(F)c(F)c(C=Cc5c(F)c(F)c(C=Cc6c(F)c(F)c(C=Cc7c(F)c(F)c(C=Cc8c(F)c(F)c(C=Cc9c(F)c(F)cc%10nsnc9%	-5.33	-3.50	1.83

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
		10)c9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12			
43	6	O=C(c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8ccc(C(=O)C(F)(F)F)cc78)c7ccc(C(=O)C(F)(F)F)cc67)c6ccc(C(=O)C(F)(F)F)cc56)c5ccc(C(=O)C(F)(F)F)cc45)c4ccc(C(=O)C(F)(F)F)cc34)scc2c1)C(F)(F)F	-5.30	-3.54	1.76
43	7	O=C(c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9ccc(C(=O)C(F)(F)F)cc89)c8ccc(C(=O)C(F)(F)F)cc78)c7ccc(C(=O)C(F)(F)F)cc67)c6ccc(C(=O)C(F)(F)F)cc56)c5ccc(C(=O)C(F)(F)F)cc45)c4ccc(C(=O)C(F)(F)F)cc34)scc2c1)C(F)(F)F	-5.27	-3.59	1.67
458	4	O=C1NC(C2=C3C(=O)NC(C4=C5C(=O)NC(C6=C7C(=O)NC=C7C(=O)O6)=C5C(=O)O4)=C3C(=O)O2)=C2C(=O)OC=C12	-5.92	-4.31	1.61
503	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sccc8C(C)=O)cc7C(C)=O)cc6C(C)=O)cc5C(C)=O)cc4C(C)=O)cc3C(C)=O)cc2C(C)=O)cc1C(C)=O	-4.92	-3.04	1.88
88	10	C=Cc1cc(C=Cc2cc(C=Cc3cc(C=Cc4cc(C=Cc5cc(C=Cc6cc(C=Cc7cc(C=Cc8cc(C=Cc9cc(C=Cc%10ccc%11ncncc%10-%11)c%10ncncc9-%10)c9ncncc8-9)c8ncncc7-8)c7ncncc6-7)c6ncncc5-6)c5ncncc4-5)c4ncncc3-4)c3ncncc2-3)c2ncncc1-2	-5.43	-4.11	1.33
98	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc8c7CCCS8(=O)=O)c7c6CCCS7(=O)=O)c6c5CCCS6(=O)=O)c	-5.08	-3.08	2.00

Monomer ID	N	SMILES	HOMO (eV)	LUMO (eV)	Gap (eV)
		5c4CCCS5(=O)=O)c4c3CCCS4(=O)=O)c3c2CCCS3(=O)=O)c2c1CCCS2(=O)=O			