Electronic Supplementary Information

A machine learning-driven prediction of Hammett constants using quantum chemical and structural descriptors

Vaneet Saini*[1], Ranjeet Kumar

Department of Chemistry & Centre for Advanced Studies in Chemistry, Panjab University, Chandigarh 160014, India. Email: vsaini@pu.ac.in

Orcid ID: 0000-0002-8186-5166

Table S1. List of molecules left out in this study

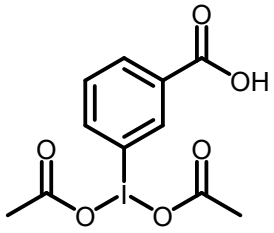
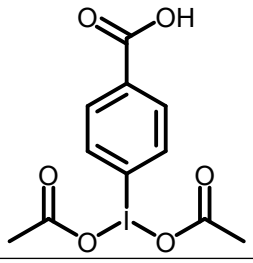
S. No	SMILES	Type	Structure	Sigma
1.	CC(=O)OI(OC(C)=O)c1cccc(C(=O)O)c1	meta		0.85
2.	CC(=O)OI(OC(C)=O)c1ccc(C(=O)O)cc1	para		0.88

Table S2. Cross-validation metrics for different models screened in this study

Model	Cross-val R^2	rmse
MLR	0.742	0.18
PLS	0.675	0.205
KNN	0.616	0.222
SVM	0.763	0.174
AR	0.714	0.191
ET	0.87	0.128
RF	0.812	0.156
ANN	0.856	0.135

Table S3. Cross-validation, train and test metrics obtained using different number of descriptors

No. of desc.	Model	Cross-val R^2	Cross-val RMSE	Train R^2	Test R^2	Test RMSE	descriptor(s)
1	ET	0.434	0.269	0.591	0.461	0.244	NBO_CC
1	NN	0.484	0.256	0.515	0.478	0.24	NBO_CC
2	ET	0.563	0.236	0.998	0.589	0.213	NBO_CC + E(HOMO)
2	NN	0.646	0.211	0.702	0.667	0.192	NBO_CC + E(HOMO)
3	ET	0.747	0.18	1	0.713	0.178	NBO_CC + E(HOMO) + IC0
3	NN	0.672	0.203	0.792	0.594	0.211	NBO_CC + E(HOMO) + IC0
4	ET	0.728	0.186	1	0.66	0.194	NBO_CC + E(HOMO) + IC0 + nBase
4	NN	0.69	0.198	0.81	0.729	0.173	NBO_CC + E(HOMO) + IC0 + nBase
5	ET	0.785	0.167	1	0.683	0.187	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z
5	NN	0.763	0.172	0.888	0.762	0.162	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z
6	ET	0.805	0.157	1	0.699	0.182	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero
6	NN	0.793	0.161	0.918	0.76	0.163	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero
7	ET	0.828	0.149	1	0.72	0.176	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD
7	NN	0.806	0.155	0.937	0.786	0.153	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD
8	ET	0.85	0.139	1	0.752	0.165	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD + NBO_SC

8	NN	0.844	0.14	0.96	0.837	0.134	NBO_CC + E(HOMO) + + IC0 + nBase + AATS3Z + nHetero + nBondsD + NBO_SC
9	ET	0.842	0.14	1	0.77	0.159	NBO_CC + E(HOMO) + IC0 +nBase + AATS3Z + nHetero + nBondsD + NBO_SC + PEOE_VSA1
9	NN	0.844	0.141	0.965	0.834	0.135	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD + NBO_SC + PEOE_VSA1
10	ET	0.855	0.136	1	0.774	0.158	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD + NBO_SC + PEOE_VSA1 + AATS7dv
10	NN	0.836	0.144	0.973	0.802	0.148	NBO_CC + E(HOMO) + IC0 + nBase + AATS3Z + nHetero + nBondsD + NBO_SC + PEOE_VSA1 + AATS7dv
20	ET	0.873	0.128	1	0.834	0.135	
20	NN	0.853	0.136	1	0.833	0.136	
50	ET	0.878	0.124	1	0.853	0.127	
50	NN	0.879	0.124	1	0.9	0.106	
283	ET	0.87	0.128	1	0.867	0.121	
283	NN	0.856	0.135	0.998	0.925	0.089	

Hyperparameter Tuning

The hyperparameter tuning was conducted using Keras Tuner, a library for systematic and efficient optimization of deep learning models. The search space included ranges for key parameters: the number of neurons in the input and hidden layers was explored using integer ranges (e.g., 32 to 128 in steps of 32), dropout rates were tested across a continuous range (0.1 to 0.5 in steps of 0.1), and the learning rate of the Adam optimizer was varied using discrete choices (1e-2, 1e-3, 1e-4). Additionally, the model dynamically adjusted the number of hidden layers, enabling the tuner to test architectures of varying depth. The build model function acted as a blueprint, constructing a neural network based on these hyperparameter combinations. Keras Tuner employed an intelligent search strategy, refining the hyperparameter space across multiple rounds (tuner/bracket=3) and selecting the best configuration after evaluating performance on the validation data. This method ensured an efficient exploration of architectures and configurations to identify the optimal parameters for the model.

The optimal hyperparameters determined through tuning define a neural network architecture and training configuration designed to balance model complexity and generalization. The first hidden layer is configured with 128 neurons and a dropout rate of 0.1, providing moderate regularization at the input stage. The second hidden layer has 64 neurons and a dropout rate of 0.1, while the third layer also has 64 neurons but with a slightly higher dropout rate of 0.3, indicating an increased need for regularization at this stage. The learning rate is set to 0.001, a typical value for the Adam optimizer, ensuring stable convergence during training. The optimal training process spans 50 epochs, starting from epoch 17 in a multi-bracket search strategy (tuner/bracket=3) with progressively refined trials (tuner/round=3), highlighting the extensive search conducted by the

tuner to identify this configuration. These parameters aim to optimize the model's performance on the task while avoiding overfitting.

The training process spans up to 200 epochs with a batch size of 32, using validation data for monitoring. The validation set was separated from the training data and constituted 10% of it. To prevent overfitting, early stopping criteria are applied with a patience of 30 epochs and a minimum improvement threshold (min_delta) of 0.001 for the monitored validation loss. Additionally, the best-performing weights from the training process are restored once early stopping is triggered. The optimal model configuration, identified after 50 epochs during a multi-bracket search strategy and refined through iterative trials, reflects a carefully tuned balance of depth, capacity, and regularization. These parameters collectively aim to maximize the model's performance while avoiding overfitting and unnecessary computational effort.

The model was trained using the new set of hyperparameters. The training process lasted for 62 epochs, after which early stopping was triggered, resulting in an improved test R^2 score of 0.935 (RMSE = 0.085), compared to 0.925 under the previous conditions. Additionally, overfitting was significantly reduced: while the training score was close to 1 with the earlier setup, it dropped to 0.971 with the new hyperparameters. These results outperform the benchmark test R^2 score of 0.930 and RMSE of 0.089 achieved by the previously developed GNN model, which utilized a graph representation of molecules. The representative metrics comparing the performance of the ET and NN models, along with the benchmark scores, are shown in Table S3.

Table S4. Hyperparameters (both tuned and untuned) for ANN and ET models, along with the performance metrics obtained from training with 283 descriptors.

S. No.	Model	Hyperparameters	Cross-val R^2	Train R^2	Test R^2	Test RMSE

1.	ANN	{
		hidden_layer1 = 64,
		dropout_layer1 = 0.1,
		hidden_layer2 = 128,
		dropout_layer1 = 0.1,
		hidden_layer3 = 128,
		dropout_layer1 = 0.1,
		optimizer = “Adam”,
		loss_function = “mse”,
		epochs = 200,
		batch_size = 32,
		lr = 0.001,
		activation function = “ReLU”
		}
			0.856	0.998	0.925	0.089

2.	ANN*	{
		hidden_layer1 = 128,
		dropout_layer1 = 0.1,
		hidden_layer2 = 64,
		dropout_layer1 = 0.1,
		hidden_layer3 = 64,
		dropout_layer1 = 0.3,
		optimizer = “Adam”,
		loss_function = “mse”,
		epochs = 62,
		batch_size = 32,
		lr = 0.001,
		activation function = “ReLU”
		
		}
			0.864	0.971	0.935	0.084

3.	ET	{bootstrap=False, max_depth=None, max_features=1.0, min_samples_leaf=1,	0.87	1.0	0.867	0.121

		min_samples_split=2, n_estimators=100, random_state = 42}				
4.	ET*	{bootstrap=False, max_depth=None, max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=600, random_state = 42}	0.783	1.0	0.860	0.124

*represents tuned parameters.

Table S5. Predicted test set values for the ANN model trained on all the 283 descriptors.

Type	SMILES	Actual	Predicted
m	O=C(O)c1cccc(B(F)F)c1	0.32	0.3
m	O=C(O)c1cccc([Ge](F)(F)F)c1	0.85	0.63
m	O=C(O)c1cccc([Si](F)(F)F)c1	0.54	0.44
m	O=C(O)c1cccc(I(F)(F)(F)F)c1	1.07	0.97
m	O=Nc1cccc(C(=O)O)c1	0.62	0.52
m	O=C(O)c1cccc(O)c1	0.12	-0.05
m	O=C(O)c1cccc(S(=O)O)c1	-0.04	0.1
m	O=C(O)c1cccc(B(O)O)c1	-0.01	0.01
m	O=C(O)c1cccc(/N=N/C(F)(F)F)c1	0.56	0.41
m	O=C(O)c1cccc(OC(F)(F)F)c1	0.38	0.41
m	O=C(O)c1cccc([Se]C(F)(F)F)c1	0.44	0.43
m	[C-]#[N+]c1cccc(C(=O)O)c1	0.48	0.44
m	N#COc1cccc(C(=O)O)c1	0.67	0.58
m	O=C(O)c1cccc(OC(F)F)c1	0.31	0.33
m	O=C(O)c1cccc(S(=O)(=O)C(F)F)c1	0.75	0.77
m	O=C(O)c1cccc(-n2nnnc2S)c1	0.45	0.43
m	O=C(O)c1cccc(OCCl)c1	0.25	0.21
m	O=C(O)c1cccc(SCF)c1	0.23	0.23
m	NC(=O)c1cccc(C(=O)O)c1	0.28	0.2
m	O=C(O)c1cccc(/C=N/O)c1	0.22	0.09
c	O=C(O)c1ccc2c(c1)OCO2	-0.16	0.08
m	NC(=S)Nc1cccc(C(=O)O)c1	0.22	0.17
m	CS(=O)c1cccc(C(=O)O)c1	0.52	0.43
m	COS(=O)c1cccc(C(=O)O)c1	0.5	0.49
m	O=C(O)c1cccc(C(=O)C(F)(F)F)c1	0.63	0.5
m	O=C(O)c1cccc(SC(F)(F)C(F)(F)F)c1	0.44	0.47
m	O=C(O)c1cccc(N(C(F)(F)F)C(F)(F)F)c1	0.4	0.48
m	O=C(O)c1cccc(S/C=C/Cl)c1	0.31	0.28
m	O=C(O)c1cccc(CC(F)(F)F)c1	0.12	0.2
m	O=C(O)c1cccc(CSC(F)(F)F)c1	0.12	0.2
m	CC(=O)c1cccc(C(=O)O)c1	0.38	0.34
m	CC(=O)Oc1cccc(C(=O)O)c1	0.39	0.34
m	COC(=O)c1cccc(C(=O)O)c1	0.37	0.34
m	CN(C)P(Cl)c1cccc(C(=O)O)c1	0.38	0.45
m	N[C+](N)SCc1cccc(C(=O)O)c1	0.13	0.25
m	CCNc1cccc(C(=O)O)c1	-0.24	-0.27
m	CS(=O)(=O)N(c1cccc(C(=O)O)c1)S(C)(=O)=O	0.47	0.49
m	O=C(O)c1cccc(S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)F)c1	0.92	0.98

m	O=C(O)c1cccc(/C=C\C(F)(F)F)c1	0.16	0.22
m	C/C=C/c1cccc(C(=O)O)c1	0.02	0.04
m	C=CCOc1cccc(C(=O)O)c1	0.09	0.1
m	CCOC(=O)c1cccc(C(=O)O)c1	0.37	0.29
m	CC(=O)OCc1cccc(C(=O)O)c1	0.04	0.17
c	Cn1c[n+](C)c2cc(C(=O)O)ccc21	1.11	0.9
m	C[Ge](C)(C)c1cccc(C(=O)O)c1	0	0.03
m	C[P+](C)(C)c1cccc(C(=O)O)c1	0.74	0.85
m	O=C(O)c1cccc([Se]C(C(F)(F)F)(C(F)(F)F)C(F)(F)F)c1	0.49	0.5
m	CC(C#N)(C#N)c1cccc(C(=O)O)c1	0.6	0.67
m	O=C(O)c1cccc(-c2ccsc2)c1	0.03	0.02
m	CC(C)(C)c1cccc(C(=O)O)c1	-0.1	0
m	CC[As](=O)(CC)c1cccc(C(=O)O)c1	0.57	0.5
m	CC[As](=S)(CC)c1cccc(C(=O)O)c1	0.52	0.45
m	CCN(CC)c1cccc(C(=O)O)c1	-0.23	-0.35
m	C[NH+](C)CCc1cccc(C(=O)O)c1	0.24	0.24
m	CC(C)(C)C(=O)c1cccc(C(=O)O)c1	0.27	0.27
m	CCC(C)(C)c1cccc(C(=O)O)c1	-0.06	-0.04
m	CCCCCc1cccc(C(=O)O)c1	-0.08	-0.07
m	O=C(O)c1cccc(-c2c(F)c(F)c(F)c(F)c2F)c1	0.26	0.34
m	O=C(O)c1cccc(-c2ccccc2)c1	0.06	0.15
m	O=C(O)c1cccc(Oc2ccccc2)c1	0.25	0.11
m	O=C(O)c1cccc(Sc2ccccc2)c1	0.23	0.2
m	CCc1ccc(-c2cccc(C(=O)O)c2)o1	0.09	0.05
m	O=C(O)c1cccc(C2CCCCC2)c1	-0.05	-0.08
m	CN(C)CCCCc1cccc(C(=O)O)c1	-0.08	-0.1
m	O=C(O)c1cccc(P(=NS(=O)(=O)C(F)(F)F)(C(F)(F)C(F)(F)C(F)(F)F)C(F)(F)C(F)(F)C(F)(F)F)c1	1.24	1.33
m	Cc1ccc(O)c(/N=N/c2cccc(C(=O)O)c2)c1	0.27	0.18
m	O=C(O)c1cccc(COc2ccccc2)c1	0.06	0.1
m	O=C(O)c1cccc(CS(=O)(=O)c2ccccc2)c1	0.15	0.29
m	O=C(O)c1cccc(/C=N/C(=O)c2ccccc2)c1	0.39	0.44
m	O=C(O)c1cccc(/C=N/NC(=O)c2ccccc2)c1	0.39	0.44
m	CCc1ccc(-c2cccc(C(=O)O)c2)cc1	0.07	0.07
m	Cc1cc(C)[n+](-c2cccc(C(=O)O)c2)c(C)c1	0.62	0.55
m	C[Si](c1ccccc1)(c1ccccc1)c1cccc(C(=O)O)c1	0.1	0.14
m	Cc1ccc(P(=O)(c2ccc(C)cc2)c2cccc(C(=O)O)c2)cc1	0.17	0.2
m	O=C(O)c1cccc([Ge](c2ccccc2)(c2ccccc2)c2ccccc2)c1	0.05	0.12
p	O=C(O)c1ccc(B(F)F)cc1	0.48	0.52
p	O=C(O)c1ccc(P(=S)(Cl)Cl)cc1	0.8	0.77
p	O=C(O)c1ccc(S(F)(F)(F)(F)F)cc1	0.68	0.78
p	O=C(O)c1ccc(I(=O)=O)cc1	0.78	0.86

p	O=C(O)c1ccc(C(F)(F)Cl)cc1	0.46	0.5
p	O=C(O)c1ccc(OC(F)(F)F)cc1	0.35	0.44
p	O=C(O)c1ccc([Se](=O)(=O)C(F)(F)F)cc1	1.21	1.08
p	[N-]=[N+]=Nc1nnnn1-c1ccc(C(=O)O)cc1	0.54	0.54
p	O=C(O)c1ccc(C(F)F)cc1	0.32	0.41
p	O=C(O)c1ccc(S(=O)(=O)C(F)F)cc1	0.86	0.9
p	O=C(O)c1ccc(SC(F)F)cc1	0.37	0.37
p	NC(=O)c1ccc(C(=O)O)cc1	0.36	0.32
p	C[Hg]c1ccc(C(=O)O)cc1	0.1	0.24
p	O=C(O)c1ccc(CO)cc1	0	-0.08
p	COSc1ccc(C(=O)O)cc1	0.17	0.22
p	CS(=O)(=O)Sc1ccc(C(=O)O)cc1	0.54	0.47
p	C[Se]c1ccc(C(=O)O)cc1	0	0.07
p	NCc1ccc(C(=O)O)cc1	-0.11	-0.13
p	O=C(O)c1ccc(N(C(=O)F)C(F)(F)F)cc1	0.5	0.25
p	O=C(O)c1ccc(P(C(F)(F)F)C(F)(F)F)cc1	0.69	0.59
p	O=C(O)c1ccc(OC(F)(F)C(F)Cl)cc1	0.28	0.33
p	O=C(O)c1ccc(/C=N/S(=O)(=O)C(F)(F)F)cc1	1	0.93
p	C#CSc1ccc(C(=O)O)cc1	0.19	0.2
p	N#CCc1ccc(C(=O)O)cc1	0.18	0.16
p	C=Cc1ccc(C(=O)O)cc1	-0.04	-0.07
p	CC(c1ccc(C(=O)O)cc1)([N+](=O)[O-])[N+](=O)[O-]	0.61	0.64
p	C=COc1ccc(C(=O)O)cc1	-0.09	-0.02
p	CCOc1ccc(C(=O)O)cc1	-0.24	-0.2
p	CCSc1ccc(C(=O)O)cc1	0.03	0.02
p	N[C+](N)SCc1ccc(C(=O)O)cc1	0.15	0.37
p	CN(C)S(=O)(=O)c1ccc(C(=O)O)cc1	0.65	0.61
p	CS(=O)(=O)N(c1ccc(C(=O)O)cc1)S(C)(=O)=O	0.49	0.52
p	CN(C)Sc1ccc(C(=O)O)cc1	0.09	0.06
p	O=C(O)c1ccc(C#CC(F)(F)F)cc1	0.51	0.43
p	O=C(O)c1ccc(S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)F)cc1	1.09	1.08
p	O=C(O)c1ccc(SC(F)(C(F)(F)F)C(F)(F)F)cc1	0.51	0.52
p	O=C(O)c1ccc(/C=C\C(F)(F)F)cc1	0.17	0.23
p	N#C/C=C/c1ccc(C(=O)O)cc1	0.17	0.27
p	CCOC(=O)c1ccc(C(=O)O)cc1	0.45	0.43
p	O=C(O)CCc1ccc(C(=O)O)cc1	-0.07	-0.1
p	CC(=O)NCc1ccc(C(=O)O)cc1	-0.05	0.01
p	CC(C)(c1ccc(C(=O)O)cc1)[N+](=O)[O-]	0.2	0.22
p	CC(O)Cc1ccc(C(=O)O)cc1	-0.17	-0.14
p	O=C(O)c1ccc(C2(F)C(F)(F)C(F)(F)C2(F)F)cc1	0.53	0.48
p	N#CC(C#N)(C#N)c1ccc(C(=O)O)cc1	0.96	1.09

p	O=C(O)c1ccc(C2(O)C(F)(F)C(F)(F)C2(F)F)cc1	0.37	0.32
p	O=C(O)c1ccc(-c2cncnc2)cc1	0.39	0.4
p	N#CC(C#N)=C(C#N)c1ccc(C(=O)O)cc1	0.98	0.97
p	O=C(O)c1ccc(-c2ccccn2)cc1	0.17	0.41
p	O=C(O)c1ccc(-c2cccnc2)cc1	0.25	0.39
p	CCCCOC(=O)Nc1ccc(C(=O)O)cc1	-0.05	-0.13
p	CCC(C)(C)c1ccc(C(=O)O)cc1	-0.18	-0.11
p	C[N+](C)(C)CCc1ccc(C(=O)O)cc1	0.13	0.17
p	O=C(O)c1ccc(P(=O)(C(F)(F)C(F)(F)C(F)(F)F)C(F)(F)C(F)(F)C(F)(F)F)cc1	1.1	1.02
p	O=C(O)c1ccc(NP(=O)(C(F)(F)C(F)(F)C(F)(F)F)C(F)(F)C(F)(F)C(F)(F)F)cc1	0.18	0.26
p	O=C(O)c1ccc(-c2ccc(Br)cc2)cc1	0.12	0.11
p	O=C(O)c1ccc(-c2cccc(F)c2)cc1	0.1	0.1
p	O=C(O)c1ccc(-c2ccc([N+](=O)[O-])cc2)cc1	0.26	0.22
p	O=C(O)c1ccc(-c2ccccc2)cc1	-0.01	0.1
p	O=C(O)c1ccc(N=Nc2ccccc2)cc1	0.39	0.46
p	O=C(O)c1ccc(OS(=O)(=O)c2ccccc2)cc1	0.33	0.36
p	O=C(O)c1ccc(Nc2ccccc2)cc1	-0.56	-0.39
p	CCc1ccc(-c2ccc(C(=O)O)cc2)o1	-0.13	-0.07
p	CN(C)CCCCc1ccc(C(=O)O)cc1	-0.16	-0.19
p	O=C(O)c1ccc(C(=O)Nc2ccccc2)cc1	0.41	0.28
p	O=C(O)c1ccc(CCc2ccccc2)cc1	-0.12	-0.16
p	CCCCP(=O)(CCCC)c1ccc(C(=O)O)cc1	0.49	0.32
p	C[Si](C)(C)O[Si](O[Si](C)(C)C)(O[Si](C)(C)C)c1ccc(C(=O)O)cc1	-0.01	-0.02
p	O=C(O)c1ccc(N(c2ccccc2)c2ccccc2)cc1	-0.22	-0.28
p	O=C(O)c1ccc(P(c2ccccc2)c2ccccc2)cc1	0.19	0.16
p	O=C(O)c1ccc(P(=S)(c2ccccc2)c2ccccc2)cc1	0.47	0.42
p	O=C(O)c1ccc(C(=O)OC(c2ccccc2)c2ccccc2)cc1	0.56	0.45

Table S6. Metrics for generalization studies.

S. No.	Seed	Total datapoints (train + test)	No. of train data points	No. of test data points (type)	Train R^2	Test R^2
1.	2	932	882	50 (meta)	0.998	0.901
2.	32	932	882	50 (meta)	0.998	0.906
3.	69	932	882	50 (meta)	0.998	0.882
4.	86	932	882	50 (meta)	0.999	0.903
5.	105	932	882	50 (meta)	0.998	0.882
		Average			**0.998**	**0.895**

Note: Default hyperparameters were used.

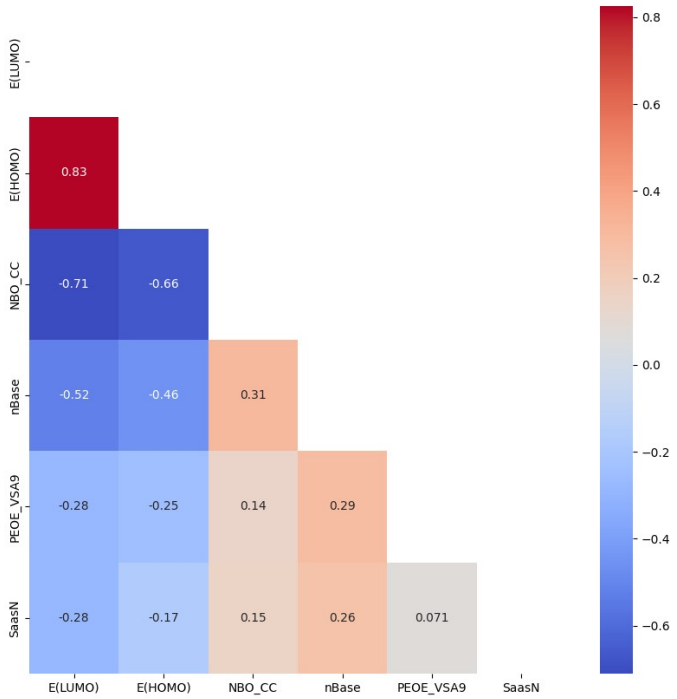


Figure S1. Correlation matric of E(LUMO) with top 5 most correlated descriptors.

Layer (type)	Output Shape	Param #
dense_13 (Dense)	(None, 128)	36,352
dropout_10 (Dropout)	(None, 128)	0
dense_14 (Dense)	(None, 64)	8,256
dropout_11 (Dropout)	(None, 64)	0
dense_15 (Dense)	(None, 64)	4,160
dropout_12 (Dropout)	(None, 64)	0
dense_16 (Dense)	(None, 1)	65

Total params: 146,501 (572.27 KB)
Trainable params: 48,833 (190.75 KB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 97,668 (381.52 KB)

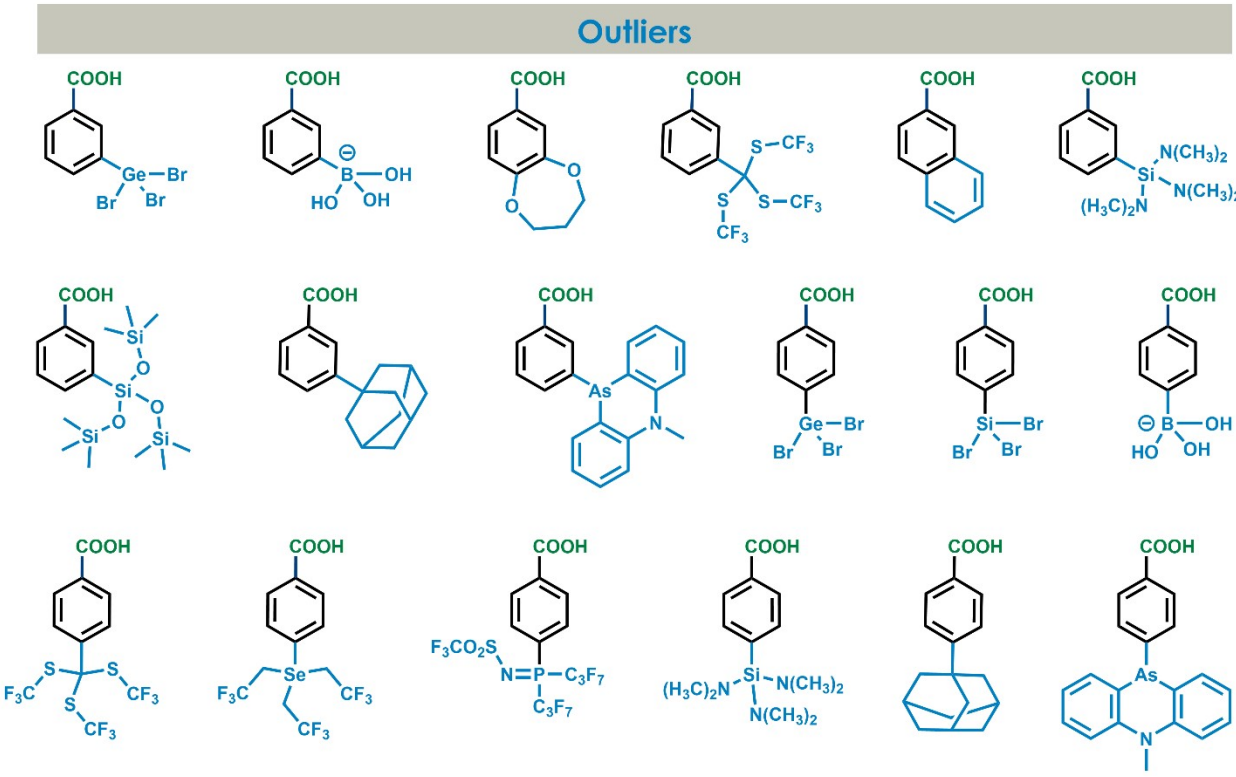


Figure S3. Outliers identified in the training set using AD analysis.