Supplementary materials

1. Definition of evaluation metrics:

The Coefficient of Determination R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 value represents the proportion of variance that has been explained by the independent variables in the model. The better predictive ability of the model, the closer to 1.0 of R^2.

The RMSE value is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}(y_i - \hat{y}_i)^2}$$

Where $\hat{y}_i$ is the predicted solubility value of the i-th compound, y_i is the experimental solubility value for the i-th compound and $\bar{y}$ is the average solubility values of n samples, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The RMSE is the square root of the mean square error which measures the differences between experimental solubility values and predicted values of a model. Thus, the closer the RMSE value is to 0, the better the model fits to the data.

2. Code for feature filtering.

```
#Filter by removing low-variance features.

from sklearn.feature_selection import VarianceThreshold

def feature_select(X,vt=0):

    selector = VarianceThreshold(vt)

    lst=selector.fit(X).get_support()

    cols=X.columns[lst]

    X_kept=X[cols]

  return X_kept

# Filter by removing high correlated descriptors. Here the input
trainx is the Pandas object of descriptors of training set

import numpy as np

import pandas as pd

trainx=pd.read_csv('train_descriptors.csv')

corr_matrix = trainx.corr().abs()

upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape),
k=1).astype(np.bool))

to_drop = [column for column in upper.columns if
any(upper[column] >= 0.90)]
```

```
to_keep=set(train.columns)-set(to_drop)

newtrain=train[list(to_keep)]
```

Table S1. List of descriptors used in model construction.

ATSe1,AW,AWeight,Arto,CIC0,CIC2,CIC3,CIC4,CIC6,Chi10,Chi3c,Chi3ch,Chi4c,Chi4ch,Chi4pc,Chi5ch,Chi6ch,Chiv10,Chiv3c,Chiv4c,Chiv4pc,Chiv5ch,Chiv6ch,EstateVSA0,EstateVSA1,EstateVSA10,EstateVSA2,EstateVSA3,EstateVSA4,EstateVSA5,EstateVSA6,EstateVSA7,EstateVSA8,EstateVSA9,GATSe1,GATSe2,GATSe3,GATSe4,GATSe5,GATSe6,GATSe7,GATSe8,GATSm1,GATSm2,GATSm3,GATSm4,GATSm5,GATSm6,GATSm7,GATSm8,GATSp1,GATSp2,GATSp3,GATSp4,GATSp5,GATSp6,GATSp7,GATSp8,GATSv1,GATSv2,GATSv3,Hy,IC0,IC1,IC2,J,LDI,LogP,LogP2,MATSe1,MATSe2,MATSe3,MATSe4,MATSe5,MATSe6,MATSe7,MATSe8,MATSm1,MATSm2,MATSm3,MATSm4,MATSm5,MATSm6,MATSm7,MATSm8,MATSp1,MATSp2,MATSp3,MATSp4,MATSp5,MATSp6,MATSp7,MATSp8,MATSv1,MRVSA0,MRVSA1,MRVSA2,MRVSA3,MRVSA4,MRVSA5,MRVSA6,MRVSA8,MRVSA9,MTPSA,Mac,Mnc,Mpc,PEOEVSA0,PEOEVSA1,PEOEVSA10,PEOEVSA11,PEOEVSA12,PEOEVSA13,PEOEVSA2,PEOEVSA3,PEOEVSA4,PEOEVSA5,PEOEVSA6,PEOEVSA7,PEOEVSA8,PEOEVSA9,QCmax,QCmin,QCss,QHmax,QHmin,QHss,QNmax,QNmin,QNss,QOmax,QOmin,QOss,Qass,Qmax,Qmin,Rnc,Rpc,S10,S11,S12,S13,S14,S15,S16,S17,S18,S19,S20,S21,S22,S23,S24,S25,S27,S28,S29,S30,S31,S32,S33,S34,S35,S36,S37,S38,S45,S46,S48,S49,S50,S51,S52,S53,S54,S7,S70,S75,S8,S9,SIC0,SIC1,SIC2,SIC3,SIC4,SIC5,SIC6,SPP,Smax10,Smax11,Smax12,Smax15,Smax16,Smax17,Smax18,Smax23,Smax27,Smax28,Smax30,Smax33,Smax34,Smax35,Smax37,Smax45,Smax53,Smax6,Smax69,Smax74,Smax8,Smin12,Smin15,Smin16,Smin8,Tnc,Tsch,UI,VSAEstate0,VSAEstate1,VSAEstate2,VSAEstate3,VSAEstate4,VSAEstate6,VSAEstate7,VSAEstate8,bcute1,bcute10,bcute11,bcute9,bcutm1,bcutm2,bcutm3,bcutm4,dchi0,dchi1,kappa3,kappam3,mChi1,naccr,naro,ndb,ndonr,nhal,nhet,nhyd,nnitro,nphos,nsb,petitjeant,slogPVSA0,slogPVSA1,slogPVSA10,slogPVSA2,slogPVSA3,slogPVSA6,slogPVSA7,slogPVSA9

Table S2. Table of descriptors in MPNN model

Feature		Description	Size
Atom Feature	atomic number	type of atom (ex. C, N, O), by atomic number	10
	Crippen atom type	Atom type defined by Crippen method and fill with corresponding Crippen atomic LogP	72
	formal charge	integer electronic charge assigned to atom	5
	# Hs	number of bonded hydrogen atoms	5
Bond Feature	bond type	single, double, triple, or aromatic	4
	conjugated	whether the bond is conjugated	1
	in ring	whether the bond is part of a ring	1
	stereo	none, any, E/Z or cis/trans	6

Table S3. Model performance in test set

Model	R2	R2_CI_Lower	R2_CI_Upper	RMSE	RMSE_CI_Lower	RMSE_CI_Upper	%LogS ± 0.7	%LogS ± 0.7_CI_Lower	%LogS ± 0.7_CI_Upper	%LogS ± 1.0	%LogS ± 1.0_CI_Lower	%LogS ± 1.0_CI_Upper

RF_Property	0.844	0.829	0.859	0.90	0.85	0.96	0.694	0.676	0.712	0.809	0.793	0.823
DNN_Property	0.845	0.830	0.859	0.90	0.86	0.95	0.687	0.669	0.705	0.814	0.799	0.829
MPNN	0.844	0.826	0.859	0.90	0.85	0.96	0.679	0.661	0.697	0.815	0.799	0.830
GraphConv	0.797	0.777	0.816	1.03	0.97	1.09	0.616	0.597	0.634	0.757	0.740	0.773
ALOGPS 2.1	0.742	0.723	0.760	1.16	1.11	1.22	0.560	0.541	0.579	0.700	0.682	0.718
ESOL Equation	0.716	0.695	0.735	1.22	1.17	1.27	0.498	0.479	0.517	0.649	0.630	0.667
RF_ESOL	0.794	0.775	0.812	1.04	0.99	1.09	0.616	0.597	0.634	0.752	0.735	0.769

Table S4. Model Performance in external test set A

Model	R2	R2_CI_Lower	R2_CI_Upper	RMSE	RMSE_CI_Lower	RMSE_CI_Upper	%LogS ± 0.7	%LogS ± 0.7_CI_Lower	%LogS ± 0.7_CI_Upper	%LogS ± 1.0	%LogS ± 1.0_CI_Lower	%LogS ± 1.0_CI_Upper
RF_Property	0.795	0.712	0.855	0.77	0.65	0.89	0.650	0.550	0.740	0.850	0.780	0.920
DNN_Property	0.780	0.685	0.844	0.80	0.70	0.90	0.640	0.540	0.730	0.770	0.680	0.850
MPNN	0.744	0.641	0.814	0.86	0.74	0.99	0.620	0.520	0.710	0.740	0.650	0.820
GraphConv	0.667	0.563	0.752	0.98	0.79	1.18	0.630	0.530	0.720	0.760	0.670	0.840
ALOGPS 2.1	0.732	0.593	0.824	0.88	0.73	1.05	0.630	0.530	0.720	0.790	0.700	0.860
ESOL Equation	0.553	0.374	0.668	1.14	0.94	1.33	0.510	0.410	0.600	0.660	0.570	0.750
RF_ESOL	0.694	0.569	0.780	0.94	0.81	1.07	0.550	0.450	0.650	0.670	0.580	0.760

Table S5. Model Performance in external test set B

Model	R2	R2_CI_Lower	R2_CI_Upper	RMSE	RMSE_CI_Lower	RMSE_CI_Upper	%LogS ± 0.7	%LogS ± 0.7_CI_Lower	%LogS ± 0.7_CI_Upper	%LogS ± 1.0	%LogS ± 1.0_CI_Lower	%LogS ± 1.0_CI_Upper
RF_Property	0.490	0.196	0.664	0.63	0.53	0.73	0.726	0.613	0.839	0.887	0.806	0.952
DNN_Property	0.507	0.188	0.696	0.62	0.51	0.73	0.790	0.694	0.887	0.887	0.806	0.952
MPNN	0.361	-0.132	0.649	0.71	0.56	0.85	0.742	0.629	0.839	0.855	0.758	0.935
GraphConv	-0.033	-0.590	0.301	0.90	0.78	1.02	0.500	0.371	0.629	0.661	0.548	0.774
ALOGPS 2.1	0.159	-0.172	0.383	0.81	0.70	0.93	0.565	0.435	0.694	0.790	0.677	0.887
ESOL Equation	0.476	0.213	0.637	0.64	0.55	0.73	0.677	0.565	0.790	0.855	0.758	0.935
RF_ESOL	0.072	-0.537	0.420	0.86	0.69	1.02	0.629	0.516	0.742	0.742	0.629	0.839

Table S6. Most relevant SHAP descriptor values of six compounds from Figure 10

ID	Exp. LogS	Pred. LogS	LogP	LogP2	Hy	bcute10	bcutm3	bcutm4	Chiv10	Tsch	PEOEVSA5
P10	-4.92	-4.36	2.564	6.572	-78.776	1.757	3.684	3.513	0.325	11318	30.332
P11	-4.82	-4.23	1.959	3.836	-75.351	1.745	3.681	3.511	0.307	11318	12.133
P12	-4.34	-3.88	2.289	5.24	-78.776	1.734	3.64	3.488	0.413	11318	12.133
P14	-3.8	-3.65	1.483	2.199	-75.351	1.721	3.635	3.479	0.349	11318	12.133
P13	-3.87	-3.72	1.768	3.127	-75.351	1.726	3.638	3.486	0.369	11318	12.133
P15	-3.65	-3.6	1.973	3.892	-70.487	1.716	3.618	3.462	0.268	9259	19.056

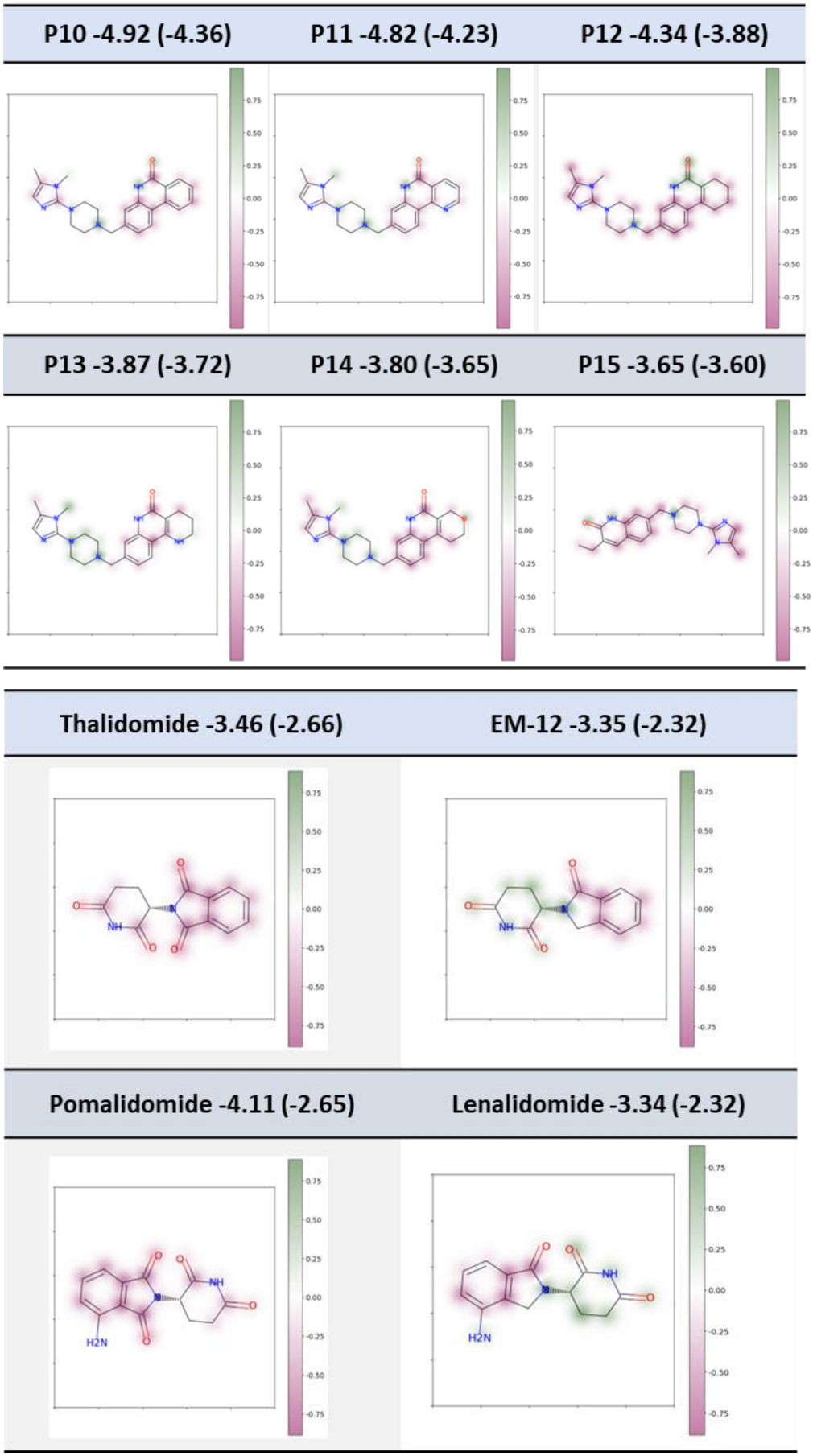
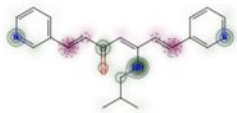
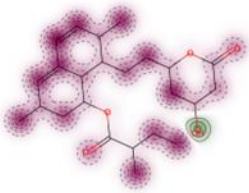
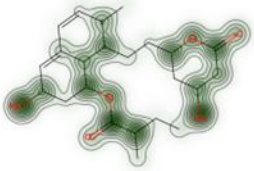
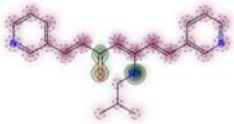
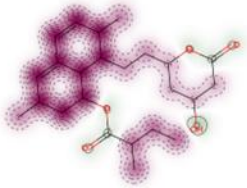
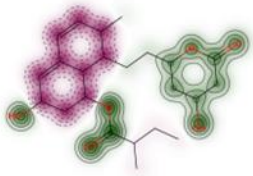
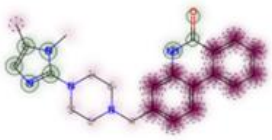
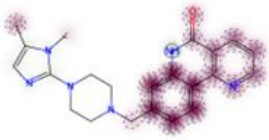
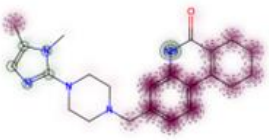
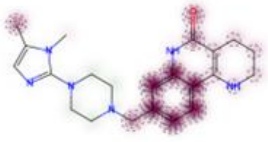
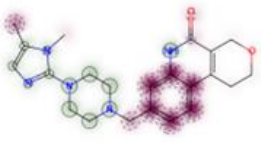
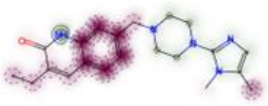


Figure S1. Interpretation results (with normalization) from atom-coloring

ID S_{exp} (S_{pred})	NC61 -3.77 (-4.20)	NC17 -3.05 (-3.35)	C-499 -6.01 (-5.75)	C1257 -3.35 (-3.67)
Atom Coloring				
Fragment Coloring				

P10 -4.92 (-4.36)	P11 -4.82 (-4.23)	P12 -4.34 (-3.88)
		
P13 -3.87 (-3.72)	**P14 -3.80 (-3.65)**	**P15 -3.65 (-3.60)**
		

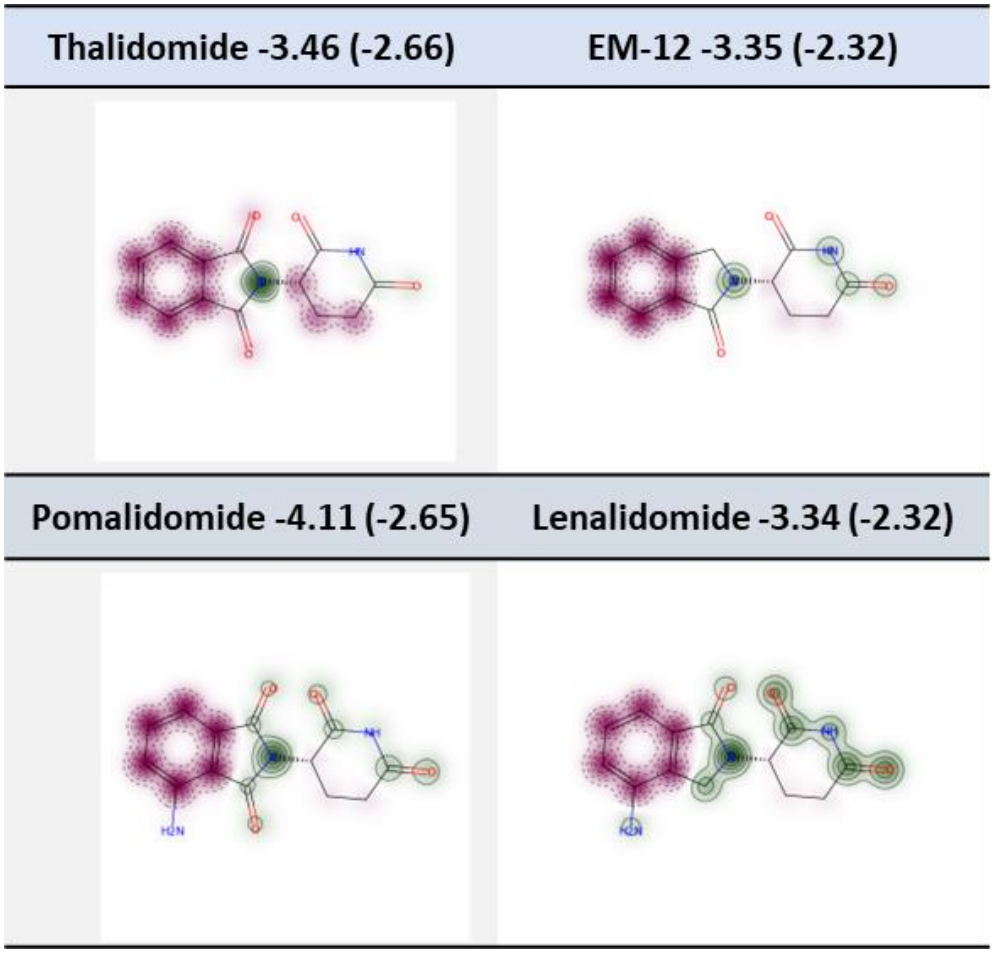


Figure S2. Interpretation results (without normalization) from atom-coloring and fragment-coloring scheme