Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction - Supporting Information

Gihan Panapitiya,* Michael Girard, Aaron Hollas, Jonathan Sepulveda, Vijayakumar Murugesan, Wei Wang, and Emily Saldanha*

Pacific Northwest National Laboratory, Richland, Washington 99352, United States

E-mail: gihan.panapitiya@pnnl.gov; emily.saldanha@pnnl.gov

Descriptors used for MDM model

Mordred

Here we list the 743 features from the Mordred package which we used for the MDM model. For more details on each descriptor category, please refer Moriwaki et al.[1] and Mordred documentation at,
https://mordred-descriptor. github.io/documentation/master/index.html.

Table S1: Mordred descriptors.

Descriptor category	Descriptor
ABCIndex	ABC, ABCGG
AcidBas	nBase, nAcid
Aromatic	nAromAtom, nAromBond
AtomCount	nI, nX, nHeavyAtom, nCl, nBr, nS, nSpiro, nH, nF, nHetero nC, nP, nB, nBridgehead, nO, nN, nAtom
Autocorrelation	ATSC3i, ATS5are, ATS0are, ATS8are, ATS3dv, ATSC1Z, ATSC5m, ATSC6dv, ATS3are, ATS7i, AATS0dv, ATSC6m, ATS6d, AATSC0pe, ATS7pe, AATS0pe, ATSC7m, ATSC0d,ATS3v, ATSC6i, ATSC5i, AATSC0v, ATS4m, ATSC1m,ATS7Z, ATS8dv, ATSC4are, ATSC4p, ATSC5d, ATSC4m,ATSC7i, ATSC5are, ATSC6v, ATS2i, ATS5pe, ATSC5dv, ATSC1are, AATS0p, ATS0dv, ATS5v, ATSC8are, ATS6are, ATSC8v, ATSC2p, ATSC0p, ATS1m, ATS0d, ATSC6Z, ATS5Z, ATS4Z, ATS2Z, ATSC3v,ATS2pe, ATS4p, ATS1v', 'ATSC8dv', 'ATS4pe', 'ATSC4i', 'ATS0m', 'ATSC2m', ATS0i, ATS1are, AATS0m, ATSC6d, ATSC8pe, ATS8m, ATSC3p, ATS5m, ATS5d, ATS5p, ATSC8d, ATSC0i, ATS1pe, ATSC6pe, ATS0pe, ATSC6are, ATSC0pe, ATSC5p,ATS7m, ATS6dv, ATS1i, AATSC0Z, ATS8i, AATSC0p, ATSC0Z, ATSC7pe, ATSC7v, ATSC1p, ATS1p, ATS1Z, ATS3Z, ATS6i, ATSC0dv, ATSC6p, ATS8p, ATSC8p, ATS4i, ATSC0are, AATS0Z, ATS3d, AATS0d, ATS2m, ATS5dv, ATS7p, ATSC2d, ATS6v, ATS1d,ATSC4pe, ATS3pe, ATS8Z, ATS1dv, ATS8d, AATSC0d, ATSC7dv, AATSC0m, ATSC0v, ATSC7d, ATSC8Z, ATSC1i, ATSC1d,ATSC5pe, ATSC1dv, ATSC2dv, ATS7dv, ATSC4dv, ATS7v, ATSC3pe, ATSC3Z, ATS3m, ATS2d, ATSC1pe, ATS0p, ATS4d, ATS8v, ATS2p, AATSC0are, AATSC0i, AATS0are, ATSC8m, ATS3p, ATS5i, ATSC4d, ATSC2v, ATS3i, ATSC2pe, ATS7are, AATS0i, ATSC2Z, ATSC4Z, ATSC3are, AATSC0dv, ATS8pe, ATS0v, ATS2dv, ATSC7Z, ATS4are, ATSC5v, ATSC7are, ATS0Z, ATS7d, ATSC1v, ATS4dv, ATSC3dv,ATS4v, ATS6pe, ATSC0m, ATSC2i, ATS6m, ATS2v, ATS2are, ATSC4v, ATS6Z, ATSC7p, ATSC3m, ATSC2are, ATSC5Z, ATSC8i, ATS6p, ATSC3d, AATS0v
BalabanJ	BalabanJ
BertzCT	BertzCT
BondCount	nBondsS, nBondsO, nBonds, nBondsA, nBondsT, nBondsKD, nBondsM, nBondsKS, nBondsD
CarbonTypes	C1SP2, C1SP1, C3SP2, FCSP3, C3SP3, C2SP1, C1SP3, C2SP3, C4SP3, C2SP2
Chi	Xp-5dv, Xc-5dv, Xpc-4dv, Xp-4dv, Xp-3d, Xch-3dv, Xpc-5dv, Xp-7d, Xc-3dv, Xp-6d, Xch-3d, Xp-6dv, Xpc-6d, Xch-4dv, Xch-5d, Xc-4d, Xp-2d, Xc-5d, Xch-7dv, Xc-3d, Xch-6dv, Xp-2dv, Xch-4d, Xch-6d, Xp-3dv, Xpc-4d, Xp-1dv,Xch-5dv, Xc-4dv, Xp-4d, Xpc-6dv, Xp-5d, Xp-7dv, Xch-7d, Xpc-5d, Xc-6dv, Xp-1d, Xc-6d
Constitutional	Spe, Sare, Si, Mpe, Mm, Sv, Mv, SZ, MZ, Mare, Sm, Mi, Sp, Mp
EStat	NssS, NsssssSn, NsssPbH, SssssSi, NsGeH3, SaaNH, NsNH3, SddsN, NsSeH, StN, SssSe, SsNH2, NaasC, NdsCH, SsCl, SdCH2, SaaS, SssdAs, NdSe, NssPH, SdSe, StsC, NaaN, NdssC, SsssCH, SssBe, NaaCH, NaaO, SsCH3, SssS, SdsCH, SssPH,SsssPbH, NsCH3, SaaCH, NsF, NssssAs SssBH, NddC, NdsssP, NssGeH2, SdssC, SdsssP, NssPbH2, SsssB, NsBr, NaaNH, SaaaC, SsssSnH, NssssBe, NssNH2, SssssP, SssssSn, NsssB, NsI, NdS, SaasC, NssBe, SsSnH3, NdssSe, NddsN, NdssS, NtN, SssCH2, NtsC, SsssNH, NsssSnH, SssssGe, NssSnH2, NsssssP, NddssSe, NsSiH3, NddssS, SaasN, SssssAs, SddssS, NaaSe, SssssBe, NssSiH2, NssAsH, NsPH2, SaaN, SsssN, SssO, SsssC, NaaaC, NsssNH, NsSnH3, SsPH2, SaaSe, SddC, NsssGe, SssssB, SssssN, SsssP, SsssGeH, SdsN, SddssSe, NdCH2, SsSeH, NsssGeH, SsGeH3, SdssSe, NssNH, SssAsH, SsLi, NdNH, SssGeH2, NsOH, SdO, NdO, NtCH, NsssssB, NssSe, SsssAs, SssPbH2, NsssssN, NsCl, SdS, SsSH, SsOH, NsssSiH, StCH, SsI, SssNH, SsSiH3,SsAsH2, NsssssPb, SdNH, NdsN, NssBH, SsPbH3, SssssSiH, SaaO, NaaS, SsF, NsAsH2, NssssdAs, NsSH, NsssssSi, SssSnH2, SssSiH2, NsPbH3,NsssP,NssCH2, NaasN, SsBr, NsssCH, SsNH3, NsNH2, NssO, SssssPb, NsssssC, SdssS, NsssN, NsLi, SssNH2, NsssAs
EccentricConnectivityIndex	ECIndex
FragmentComplexity	fragCpx
Framework	fMF

Table S2: Mordred descriptors continued.

Descriptor category	Descriptor
HydrogenBon	nHBDon, nHBAcc
InformationContent	CIC5, TIC2, ZMIC3, IC5, ZMIC1, MIC5, ZMIC5, IC0, TIC3, MIC3, ZMIC2, TIC5, MIC0, CIC1,CIC0, MIC4, ZMIC0,TIC0, IC2, IC3,TIC1, MIC2, IC4, CIC4, TIC4, CIC3, ZMIC4, MIC1, IC1, CIC2
Lipinski	Lipinski, GhoseFilter
LogS	FilterItLogS
McGowanVolu	VMcGowan
MoeTyp	VSA_EState2, SlogP_VSA11, SlogP_VSA9, SMR_VSA6, EState_VSA6, PEOE_VSA13, SlogP_VSA6, PEOE_VSA9, VSA_EState7, SMR_VSA7, SlogP_VSA2, VSA_EState6, SlogP_VSA8, PEOE_VSA6, EState_VSA9, EState_VSA1, VSA_EState8, EState_VSA3, SlogP_VSA5, EState_VSA4, VSA_EState9, VSA_EState1, PEOE_VSA1, SMR_VSA5, SMR_VSA8, SlogP_VSA10, PEOE_VSA7, VSA_EState4, SlogP_VSA7, VSA_EState5, EState_VSA8, SMR_VSA2,PEOE_VSA10, PEOE_VSA3, LabuteASA, PEOE_VSA8, SMR_VSA3, EState_VSA5, EState_VSA10, SlogP_VSA1, PEOE_VSA2, SMR_VSA4, PEOE_VSA11, VSA_EState3, PEOE_VSA4, SMR_VSA1, PEOE_VSA5, EState_VSA7, SlogP_VSA4, SlogP_VSA3, PEOE_VSA12, SMR_VSA9, EState_VSA2
PathCount	piPC1, TMPC10, piPC10, piPC3, MPC10, MPC9, MPC8, MPC5, piPC6, piPC5, piPC8, piPC9, piPC7, MPC2, piPC2, MPC7, TpiPC10, piPC4, MPC6, MPC4, MPC3
Polarizability	bpol, apol
RingCount	n9FAHRing, n5aRing, nG12FHRing, n5HRing, n11AHRing, n5AHRing, n3HRing, n12AHRing, n9ARing, n6FHRing, nARing, n4HRing, n10aRing, n9Ring, n5FaHRing, nRing, naRing, n6FaRing, n10HRing, n6FaHRing, n6ARing, nFaRing, n9HRing, n6aRing, n5FARing, n8FARing, n4FaRing, nG12Ring, n4ARing, n11FRing, n4AHRing, n12aRing, n12FRing, n9aRing, n12FARing, n5FaRing, n3AHRing, naHRing, n9FaHRing, n5FRing, n6aHRing, n12FaRing, n11FHRing, n8FHRing, n7FaRing, n3aRing, n7FARing, n8FaHRing, n4FARing, n6FAHRing, n5FHRing, n12FaHRing, n3aHRing, n9aHRing, nFARing, nFaHRing, n6Ring, n11FaRing, nG12FaHRing, n9FARing, nFAHRing, n12FHRing, n12HRing, n12Ring, n10aHRing, n4aRing, nG12FARing, n3Ring, n10FaHRing, n8FRing, n9FRing, nAHRing, n10AHRing, n9FHRing, n7AHRing, n8aHRing, nFHRing, n12ARing, n12aHRing, n9AHRing, nG12FAHRing, n8FAHRing, nG12AHRing, nG12FRing,n11FARing, n8ARing, n8FaRing, n4aHRing, n4FHRing, n11HRing, n10Ring, n10FRing, n4FRing, n7HRing, nHRing, n5FAHRing, n7FRing, n6HRing, n12FAHRing, n7aHRing, n10FaRing, n4FaHRing, n11aRing, nG12ARing, n10FARing, n5Ring, n3ARing, n10ARing, n11FAHRing, n11FaHRing, n9FaRing, n7FAHRing,nG12FaRing, n6FRing, n10FAHRing,n8AHRing, nG12aRing, nFRing, n5aHRing, n11Ring, n6FARing, nG12aHRing, nG12HRing, n4FAHRing, n7FHRing, n7aRing, n11aHRing, n11ARing, n7ARing, n8HRing, n4Ring, n6AHRing, n5ARing, n8Ring, n10FHRing, n8aRing, n7FaHRing, n7Ring
RotatableBon	[nRot]
SLogP	[SLogP, SMR]
TopoPSA	TopoPSA, TopoPSA(NO)
TopologicalCharg	GGI8, JGI7, GGI10, JGI5, JGI6, JGI1, JGI3, JGI8, JGI2, GGI9, GGI5, GGI6, GGI4, JGI4, JGI9, JGT10, JGI10, GGI2, GGI1, GGI3, GGI7
TopologicalIndex	Diameter, Radius
WalkCount	MWC02, MWC03, TSRW10, SRW03, SRW08, MWC04, SRW04, MWC06, TMWC10, SRW07, MWC08, SRW10, MWC01, MWC10, SRW09, SRW02,MWC05, MWC09, SRW06, MWC07, SRW05
Weight	AMW, MW
WienerIndex	WPol, WPath
ZagrebIndex	Zagreb1, mZagreb2, Zagreb2

Fragments descriptors

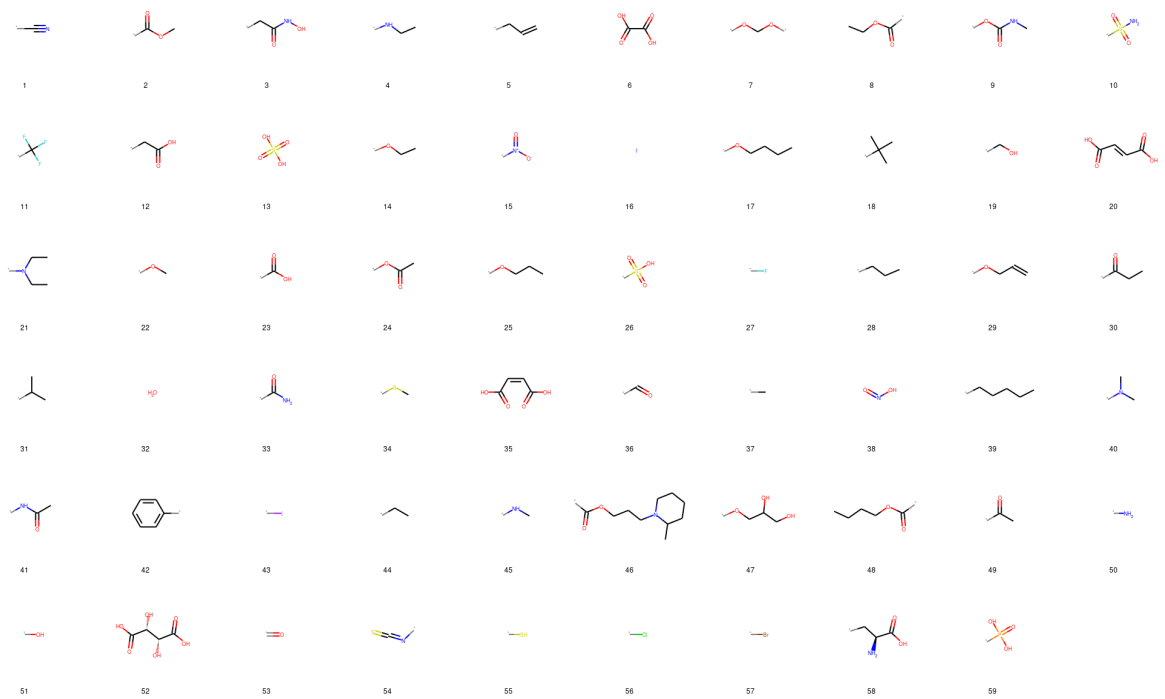


Figure S1: Most common fragments determined using RDKit (1-52). Other commonly found fragments identified with the help from a chemistry expert (53-59).

Comparison of structural properties of different datasets

We present further comparison of the properties of our dataset with external datasets used for performance comparison.

Table S3: Additional structural properties of molecules in the datasets used in this work. H.Atom and A.Bond are the counts of heavy atoms atoms and aromatic bonds. *Cl,*C,*=O, and *O are the counts of fragments containing -Cl, -C, =O and -OH, where "*" denotes any arbitrary atom.

Dataset	Mass	H.Atom	A.Bond	*Cl	*C	*=O	*O
Ours	16 - 1817	1 - 132	0 - 66	0 - 12	0 - 24	0 - 17	0 - 33
Delaney	16 - 780	1 - 55	0 - 30	0 - 12	0 - 7	0 - 6	0 - 11
Tang	46 - 665	2 - 47	0 - 27	0 - 10	0 - 7	0 - 6	0 - 8
Cui	16 - 1582	1 - 109	0 - 64	0 - 12	0 - 24	0 - 17	0 - 20
Boobier	111 - 460	7 - 33	0 - 24	0 - 3	0 - 4	0 - 4	0 - 6
Huuskonen	46 - 665	2 - 47	0 - 27	0 - 10	0 - 7	0 - 6	0 - 6
Sol. Challenge 1	138 - 504	8 - 36	0 - 21	0 - 4	0 - 7	0 - 4	0 - 6
Sol. Challenge 2 SET1	152 - 1201	11 - 85	0 - 27	0 - 2	0 - 24	0 - 11	0 - 5
Sol. Challenge 2 SET2	151 - 846	11 - 61	0 - 32	0 - 4	0 - 11	0 - 4	0 - 3
Water Set Wide (WSW)	26 - 494	2 - 35	0 - 27	0 - 10	0 - 7	0 - 5	0 - 8
Water Set Narrow (WSN)	26 - 494	2 - 33	0 - 19	0 - 4	0 - 6	0 - 5	0 - 8
Hou SET1	165 - 405	12 - 25	0 - 18	0 - 8	0 - 5	0 - 3	0 - 3
Hou SET2	70 - 459	4 - 24	0 - 21	0 - 9	0 - 3	0 - 3	0 - 4
Wang	16 - 780	1 - 55	0 - 30	0 - 12	0 - 7	0 - 6	0 - 11

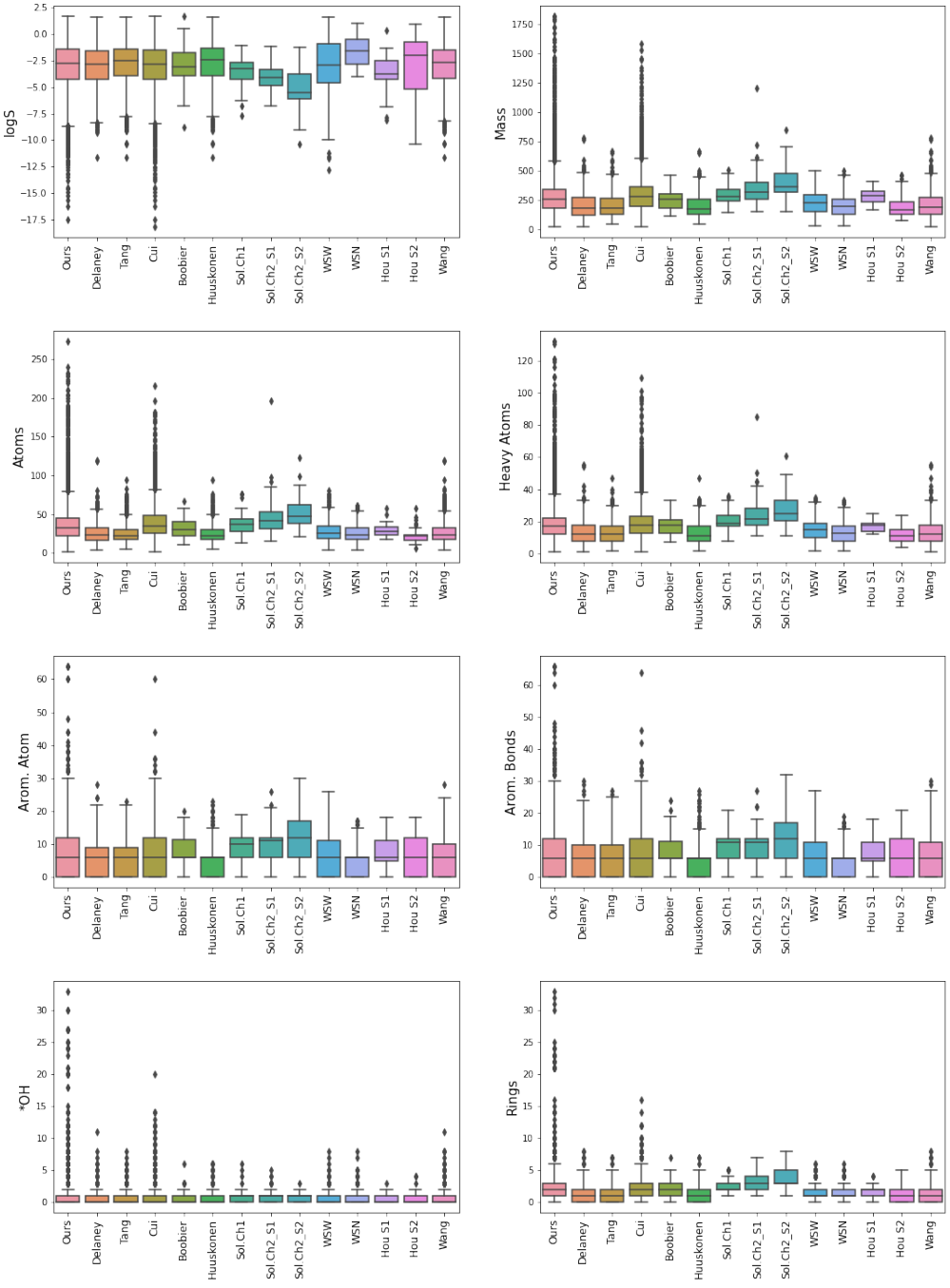


Figure S2: Box plots showing the distribution of different structural properties of solubility datasets. Sol.Ch1, Sol.Ch2_S1, and Sol.Ch2_S2 stand for Sol. Challenge 1, Sol. Challenge 2 SET1, and Sol. Challenge 2 SET2 respectively.

Duplicate removal process

Because we perform analysis of the model performance on different combinations of our data with external datasets, we must deal with duplicate entries that exist across the datasets. We used a process similar to what is described in[2] to resolve duplicates in the external datasets. If the number of duplicates is exactly two and the difference between these solubilities is less than 0.03, two entries were merged by considering the average of the two values. If the difference is greater than 0.03, the values were discarded. If the number of duplicates are greater than two, their standard deviation was calculated. If the standard deviation is less than 0.05, the solubility closest to the mean solubility of the duplicates was kept in the dataset and the other were discarded. Next, the SMILES were converted to the format defined in RDKit in order to make sure that all the SMILES strings across all the datasets conform to the same convention. If any duplicate SMILES resulted after this conversion, all such duplicates were discarded as the duplication might have been caused by limitations in RDKit. This process discarded 22, 2 and 8 molecules in Delaney, Huuskonen and Solubility Challenge 1 datasets respectively. In the Delaney dataset, additional 11 pairs of SMILES strings which were detected to be duplicates after being read by RDKit, were also removed. In the Huuskonen dataset, 4 SMILES strings failed to be read by RDKit and another 2 pairs of SMILES happened to be duplicates after converting to the canonical form.

Structure-Solubility Exploration

Before applying predictive modeling to our dataset, we first explore the data and perform a structure-property relationship study. We start by analyzing the relationships between measured solubility values and molecular descriptors as calculated by Mordred. Features that correlate with log solubility with an absolute Pearson correlation coefficient greater than 0.4 are shown in Figure S3. Positive and negative correlations are indicated using blue and red colors respectively. This result only includes features that do not contain

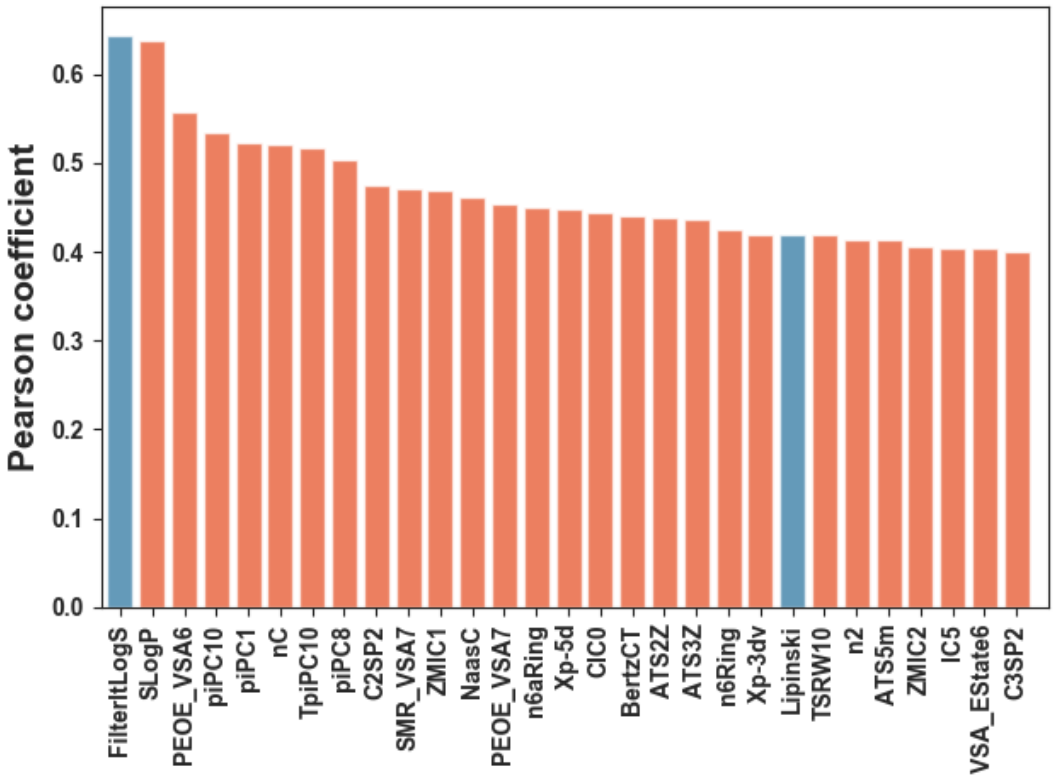


Figure S3: Absolute value of Pearson correlations of highly correlated features with log solubility with positive correlations in blue and negative correlations in red

any missing values. Additionally, because we expect many of the molecular descriptors to be highly correlated, we also removed any features with a correlation coefficient greater than 0.95 with another feature while keeping the feature that has higher correlation with solubility. This allows us to identify features which likely provide independent predictive signals of solubility. Pairwise scatter plots of most positively and negatively correlated descriptors are shown in Figure S5 and Figure S6. We next describe some of the 29 molecular descriptors identified to have an absolute correlation greater than 0.4.

FilterItLogS is a theoretical approximation for solubility originally used in the Filter-it software, which explains why this feature has such high observed correlation with solubility. *FilterItLogS* is calculated as,

$$\mathrm{FilterItLogS} = 0.898 + 0.104\sqrt{\mathrm{MOLWT}} + w_i c_i, \tag{S1}$$

where MOLWT is the molecular weight, and w_i is the weight corresponding to the

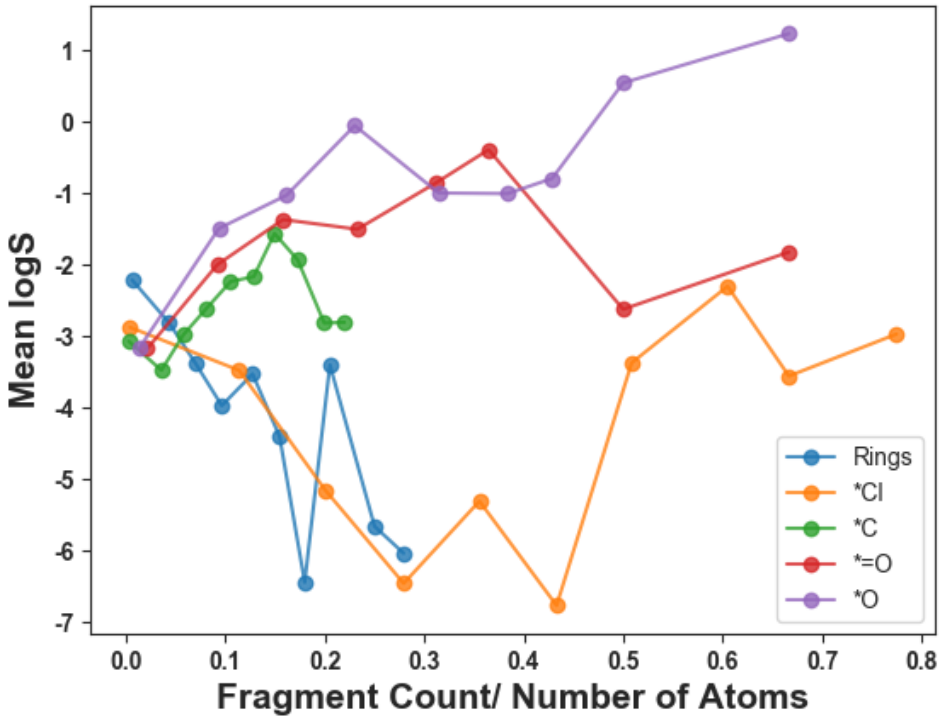


Figure S4: Mean solubility of molecules within different bins of the number of given fragments and rings in the molecules normalized by molecular size.

count of i^{th} fragment c_i. The types of fragments used are listed in Table S4 along with the corresponding weights used in the RDKit implementation of *FilterItLogS*.

In general, the fragments containing N and O have positive weights and those containing C and halogen atoms have negative weights. Therefore, the high correlation of *FilterItLogS* with solubility implies that the solubility is proportional to the counts of O and N containing fragments and inversely proportional to the molecular weight and the number of halogens in the molecule.

SlogP is the octanol-water partition coefficient calculated based on the method proposed by Wildman and Crippen.[3] This correlation is also not surprising as the octanol-water partition coefficient has been known to correlate with solubility.[4,5] *PEOE_VSA6*, *SMR_VSA7* and *PEOE_VSA7* are measures of van der Waals surface area of the atoms. The fact that these terms having negative correlations with solubility indicates that the larger the size of the molecule, the less soluble it is. In fact, the molecular size is consid-

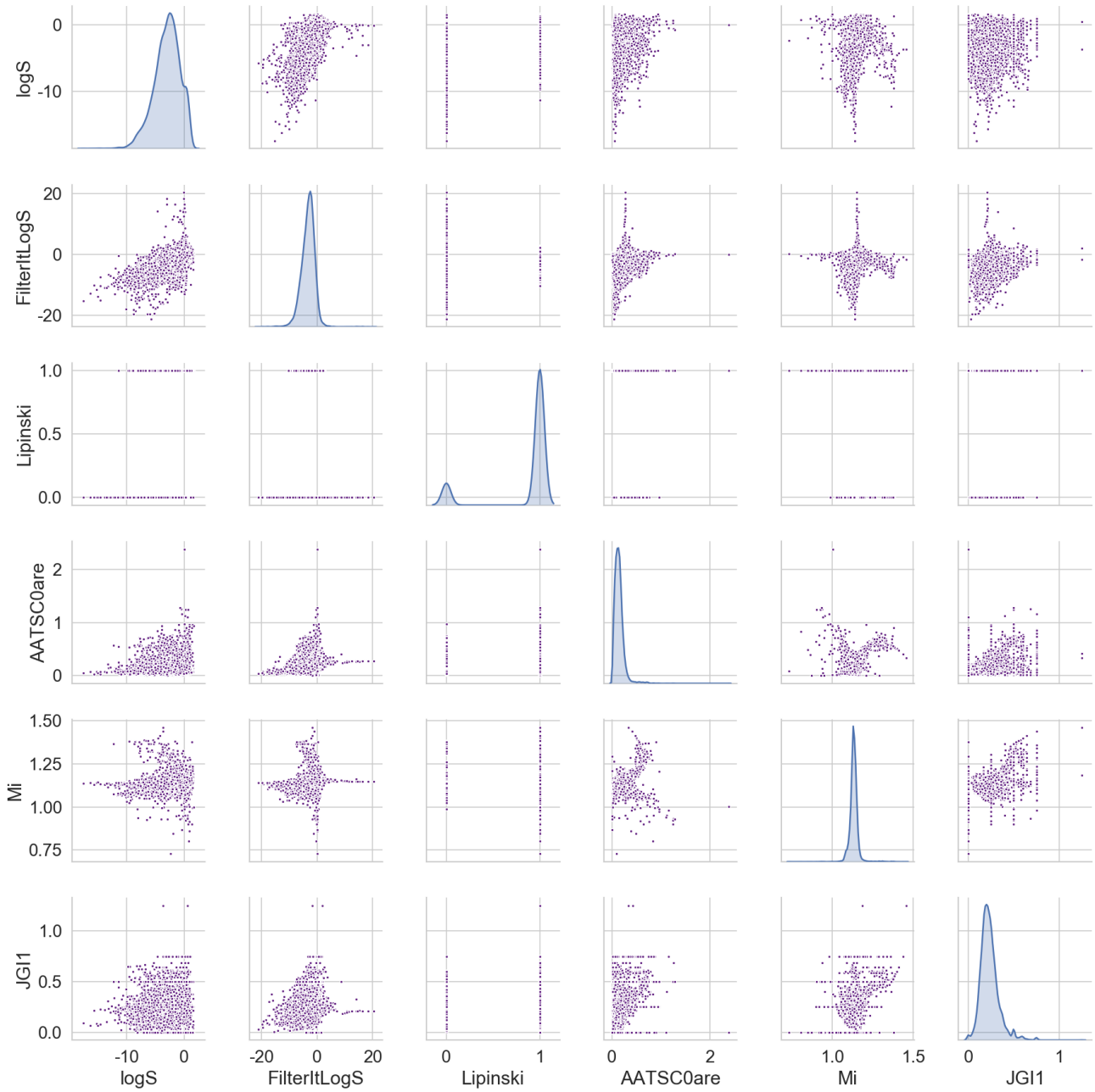


Figure S5: Top 5 positively correlated features.

ered as an important feature for solubility prediction.[6] *piPC1*, *piPC8*, and *piPC10* are path counts (that are weighted by bond order) of length 1, 8 and 10 respectively. *TpiPC10* is sum of weighted path counts over the path lengths 1 to 10. *nC* is the number of Carbon atoms. *C2SP2* is the number of SP2 carbon atoms bound to two other carbons. *ZMIC1* is a measure of the information content (Shannon's entropy) calculated by classifying atoms based on the bond order and the type of the neighboring atom. *NaasC* is the number of

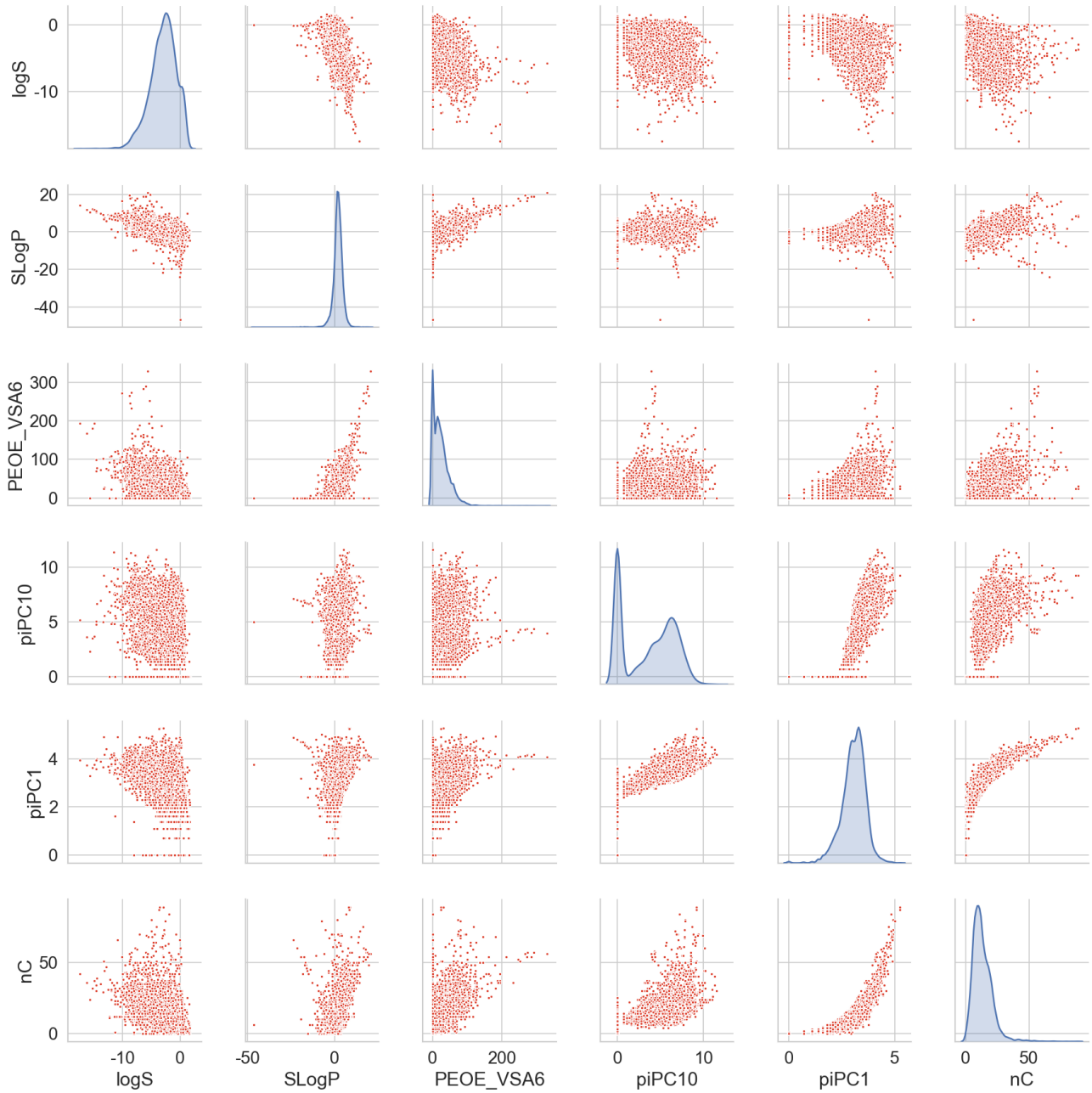


Figure S6: Top 5 negatively correlated features.

carbon atoms to which two aromatic bonds and a single bond are attached. *n6aRing* is the number of 6-membered aromatic rings in the molecule. *Xp-5d* is the Chi connectivity index (weighted by sigma electrons) for fragments containing 5 bonds. *CIC0* is the complementary information content based on different types of atoms in the molecule. *BertzCT* is a measure of “complexity" of a molecule. According to RDKit’s documentation, this feature consists of two parts to quantify the complexity of bonding and the

Table S4: Fragments and weights used for FilterItLogS

Fragment	**Weight**
$[NH0; X3; v3]$	0.71535
$[NH2; X3; v3]$	0.41056
$[nH0; X3]$	0.82535
$[OH0; X2; v2]$	0.31464
$[OH0; X1; v2]$	0.14787
$[OH1; X2; v2]$	0.62998
$[CH2; !R]$	-0.35634
$[CH3; !R]$	-0.33888
$[CH0; R]$	-0.21912
$[CH2; R]$	-0.23057
$[ch0]$	-0.37570
$[ch1]$	-0.22435
F	-0.21728
Cl	-0.49721
Br	-0.57982
I	-0.51547

distribution of heteroatoms.

In the Mordred implementation, the *Lipinski* value for a given molecule is a binary indicator of whether all four of Lipinski rules are satisfied (number of Hydrogen bond donors $\leq$ 5, number of Hydrogen bond acceptors $\leq$ 10, molecular weight $\leq$ 500, and logP $\leq$ 5). According to Lipinski, an orally active drug should satisfy at least three of these rules. Therefore, some information regarding solubility should be embedded in these rules. However, as seen from Figure S3 many molecules with high measured solubility have Lipinski values of zero.

Another way to explore the trends in solubility is in terms of molecular fragments. Early work on solubility often used the counts and fractions of fragments as a predictive signal. As we saw with the *FilterItLogS* descriptor, such features have high correlation with the molecular solubility.

We trained a random forest model using the log solubility as the target property and the counts of 59 fragments (described in the Data section of the main text) within the molecule as the features. Cl, C, =O, and O are among the highly influential features according to random forest's feature importance metric (the feature importance

scores of the most important ten features are given in Figure S7). In Figure S4, we show how the solubility changes with respect the prevalence of these four fragments within a molecule. Since the raw fragment counts are likely to be correlated with the size of the molecule (which also affects the solubility), we normalize the molecular fragment counts by the total size of molecule in terms of number of atoms. Not surprisingly, O containing fragments have positive impact on the solubility; the higher the number O-containing fragments the higher the solubility. The O containing fragments are instrumental in forming hydrogen bonds with water during the solvation process. C denotes C-A single bonds, where A can be any atom. The presence of single bonds involving C appears to be favourable for solubility, when at most the count of such bonds is less than 20% of the number of atoms in the molecule. Cl containing fragments seem to have mixed effects on solubility. As long as the number of Cl atoms constitute less than 40% of the total number of atoms (assuming that Cl-Cl fragments do not exist), an increase in the Cl content of the molecules results in a reduction in solubility. This is expected as halogen bonds are known to be hydrophobic.[7] However, when the Cl content increases further we see the the solubility of the corresponding molecules increase. This is probably due to the effect of other atoms and functional groups in the molecule.

As 80% of the molecules in our dataset contain rings, it is also interesting to analyze how solubility is related to the number of rings. In addition to the specified fragments discussed above, Figure S4 also shows the relationship between solubility and the number of rings relative to the molecules size. We find that molecules that have higher numbers of rings relative to their size tend to have reduced solubility.

In addition to the presence of different molecular fragments, the position and structure of fragments/functional groups can also have a significant effect on the solubility. As we will show later in the text, predicting the solubility of positional isomers is a challenging task. In order to understand the level of solubility variation among groups of similar molecules, we considered three sets of molecules: (1) positional isomers, (2) molecules with same core structures but different functional groups, and (3) molecules containing same number and type of functional groups attached to different core structures. For

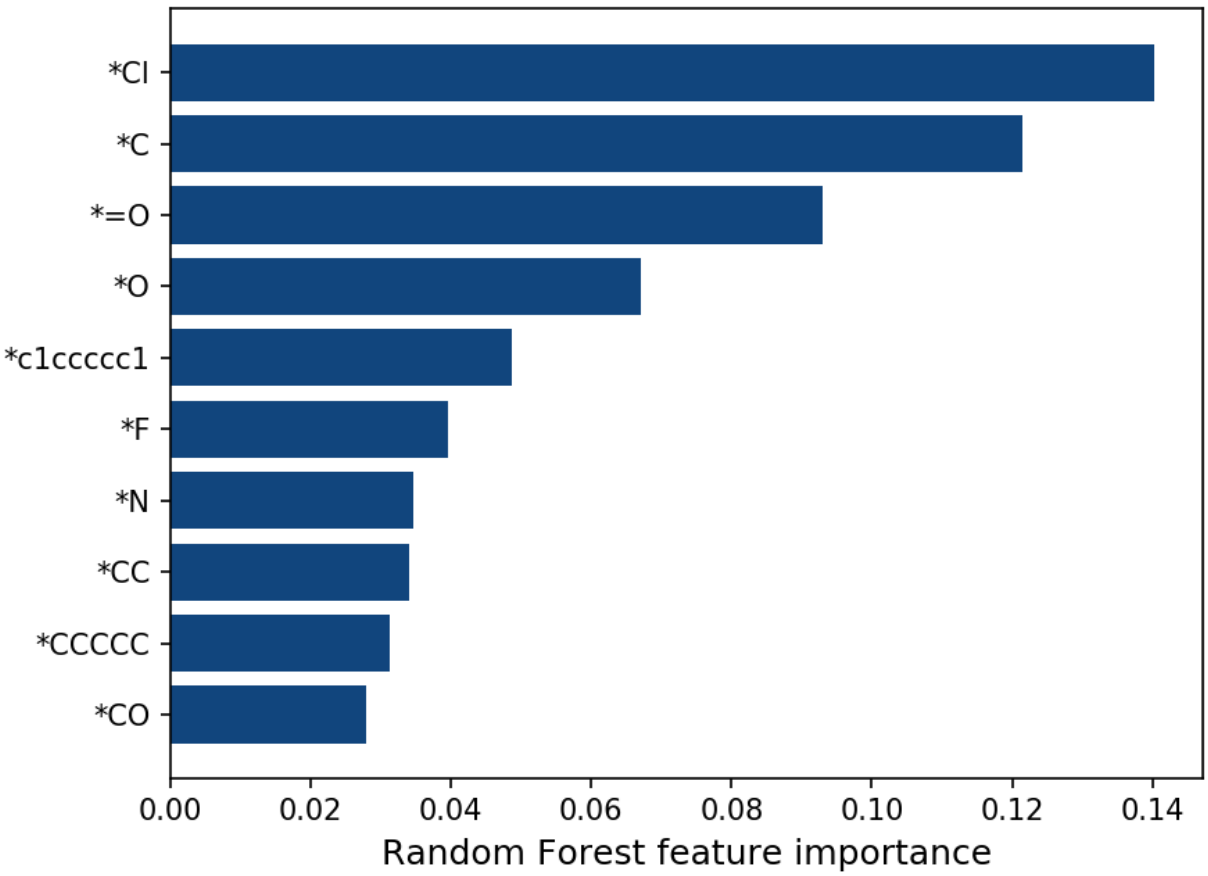


Figure S7: Feature importance of top 10 highly important fragments as determined by the random forest algorithm.

example, there are 468 groups of molecules in the isomer set, where each such group consists of n molecules that are isomers of each other. Correspondingly, there are 176 groups of molecules with the same core structure (we excluded isomers from this set) and 21 groups of molecules having the same number and type of functional groups but different core structures. The median number of molecules in isomer, same-core and same-functional-group sets are 2, 4 and 37 respectively. In Figure S8 we compare the level of solubility variability that exists in these groups relative the level of variability in random groups of molecules of the same size. We see that in all cases the groups of similar molecules had less variation in solubility than random groups of molecules. However, the level of variability among the isomer groups was significantly lower than the other two group types.

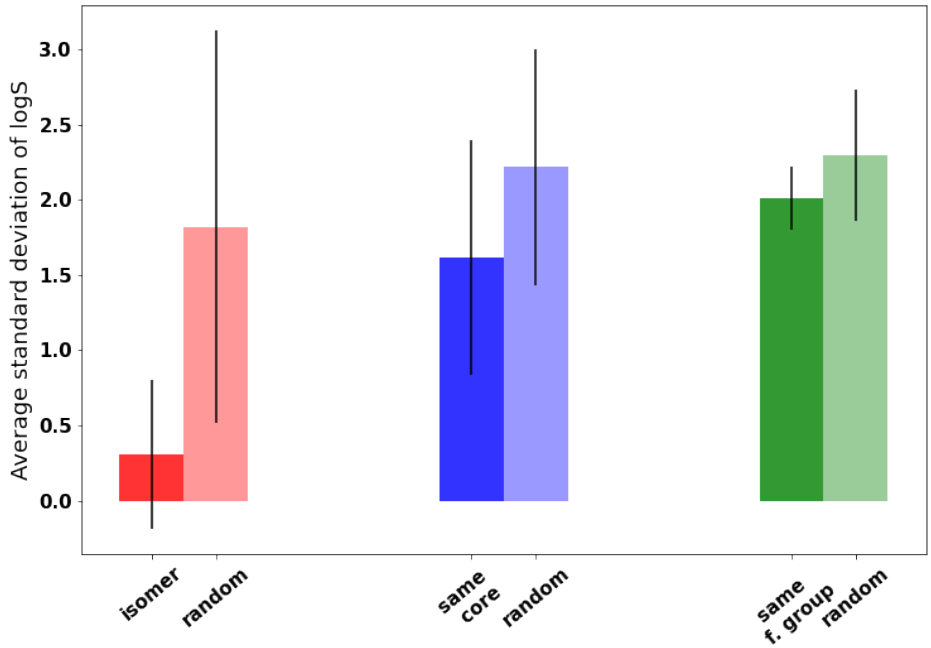


Figure S8: Average standard deviation of solubilities of molecules in sets of isomer, same core, and same functional groups.

Graph Neural Network Architecture

The graph neural network architecture leveraged in this work uses an iterative process called message passing to update the node and edges features during training. At each iteration, node features of node i ($\mathbf{x}_i$) are updated according to, $\mathbf{x}_i^t = \gamma^{t-1}(\mathbf{x}_i^{t-1}, m_i^{t-1})$, where γ^{t-1} is the update function which is a differentiable function like a multi layer perception and m_i^{t-1} is the aggregated message coming from the neighboring nodes given by

$$m_i^{t-1} = \Lambda_{j \in \mathcal{N}(i)} \, \phi_{\mathbf{\Theta}} \left(\mathbf{x}_i, \mathbf{x}_j, \mathbf{e}_{j,i} \right). \tag{S2}$$

$\mathcal{N}(i)$ are the neighboring atoms to atom i. Λ is a differentiable function that is used to aggregate the message of a given node with those of its neighboring ones. Usually, this function is one of summation, mean, or max. $\phi_{\mathbf{\Theta}}$ is another differentiable function like a multi layer perceptron.

For graph convolution networks (GCNs), m_i^{t-1} takes the form, $\sum_{j \in N(i)} \frac{1}{\sqrt{\deg(i)}} \frac{1}{\sqrt{\deg(j)}} \Theta \cdot$

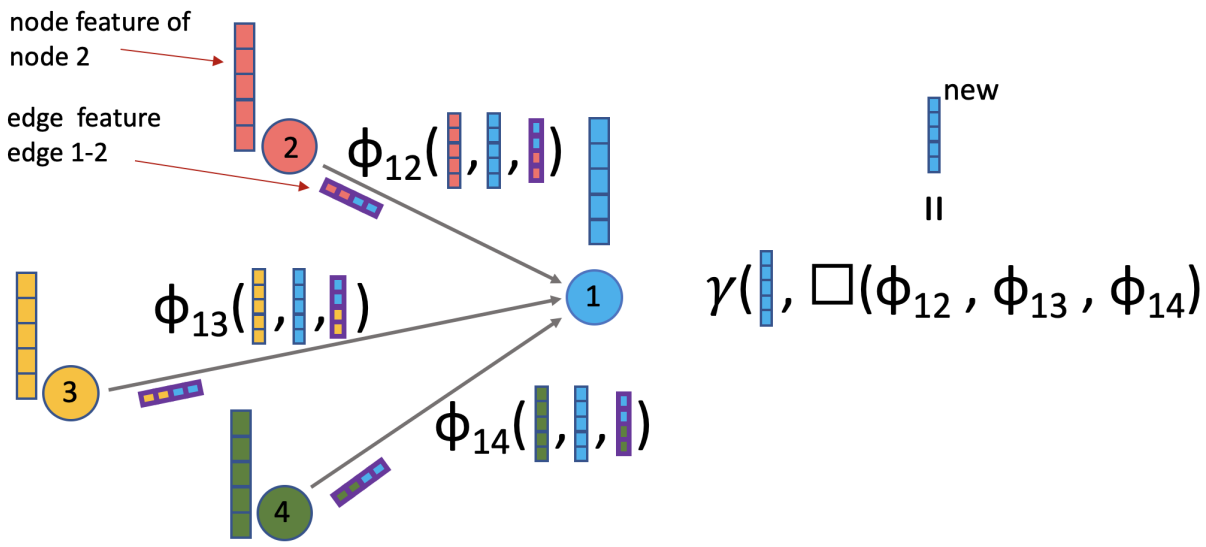


Figure S9: A depiction of how message passing works in a graph neural network.

x_i^{t-1}, and the update function is summation.[8,9] Thus the node features in a GCN are updated as,

$$x_i^t = \sum_{j \in N(i)} \frac{1}{\sqrt{\deg(i)}} \frac{1}{\sqrt{\deg(j)}} (\Theta \cdot x_j^{t-1}), \tag{S3}$$

where Θ is a weight matrix used to linearly transform the node features and $\deg(i)$ is the degree of the i^{th} node.

We also use an edge convolutional layer[10] for which the edge representations are updated according to,

$$\mathbf{x}_i^t = \sum_{j \in \mathcal{N}(i)} h_{\mathbf{\Theta}}(\mathbf{x}_i^{t-1} \,\|\, \mathbf{x}_j^{t-1} - \mathbf{x}_i^{t-1}). \tag{S4}$$

Here $h_{\mathbf{\Theta}}$ represents an arbitrary neural network and $||$ denotes concatenation of two vectors. An illustrated description on the general mechanism of message passing and aggregation in a graph neural network is shown in Figure S6.

For the GNN considered in this work our node feature vector consists of 65 elements.

1. Atomic symbol (as a one hot encoded vector of [Ag, Al, As, B, Br, C, Ca, Cd, Cl, Cu, F, Fe, Ge, H , Hg , I, K, Li, Mg, Mn, N, Na, O, P, Pb, Pt, S, Se, Si, Sn, Sr, Tl , Zn, Unknown]

2. Degree of the atom (as a one hot encoded vector of [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
3. Implicit valence of the atom (as a one hot encoded vector of [0, 1, 2, 3, 4, 5, 6])
4. Formal charge
5. Number of radical electrons
6. Hybridization of the atom (as a one hot encoded vector of [SP, SP2, SP3, SP3D, SP3D2])
7. Is the atom aromatic? (Boolean value)
8. Total number of hydrogen atoms (as a one hot encoded vector of [0, 1, 2, 3, 4])

Binning solubilities for stratified splitting of the database into train/test and validation folds

We sample our train and test sets using a stratified sampling approach based on solubility values.

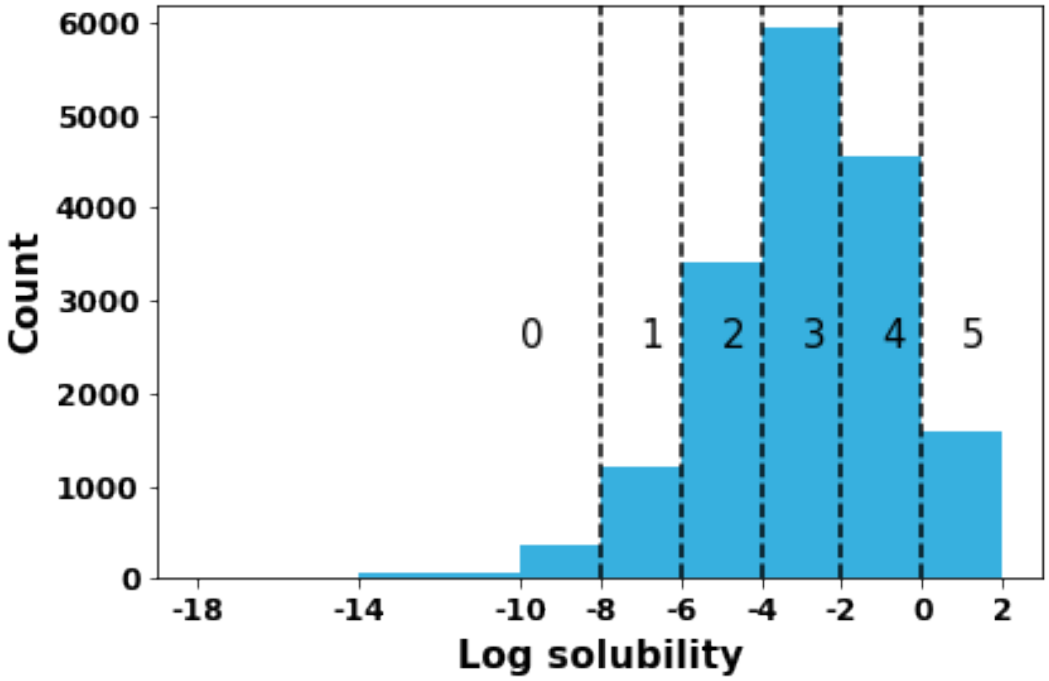


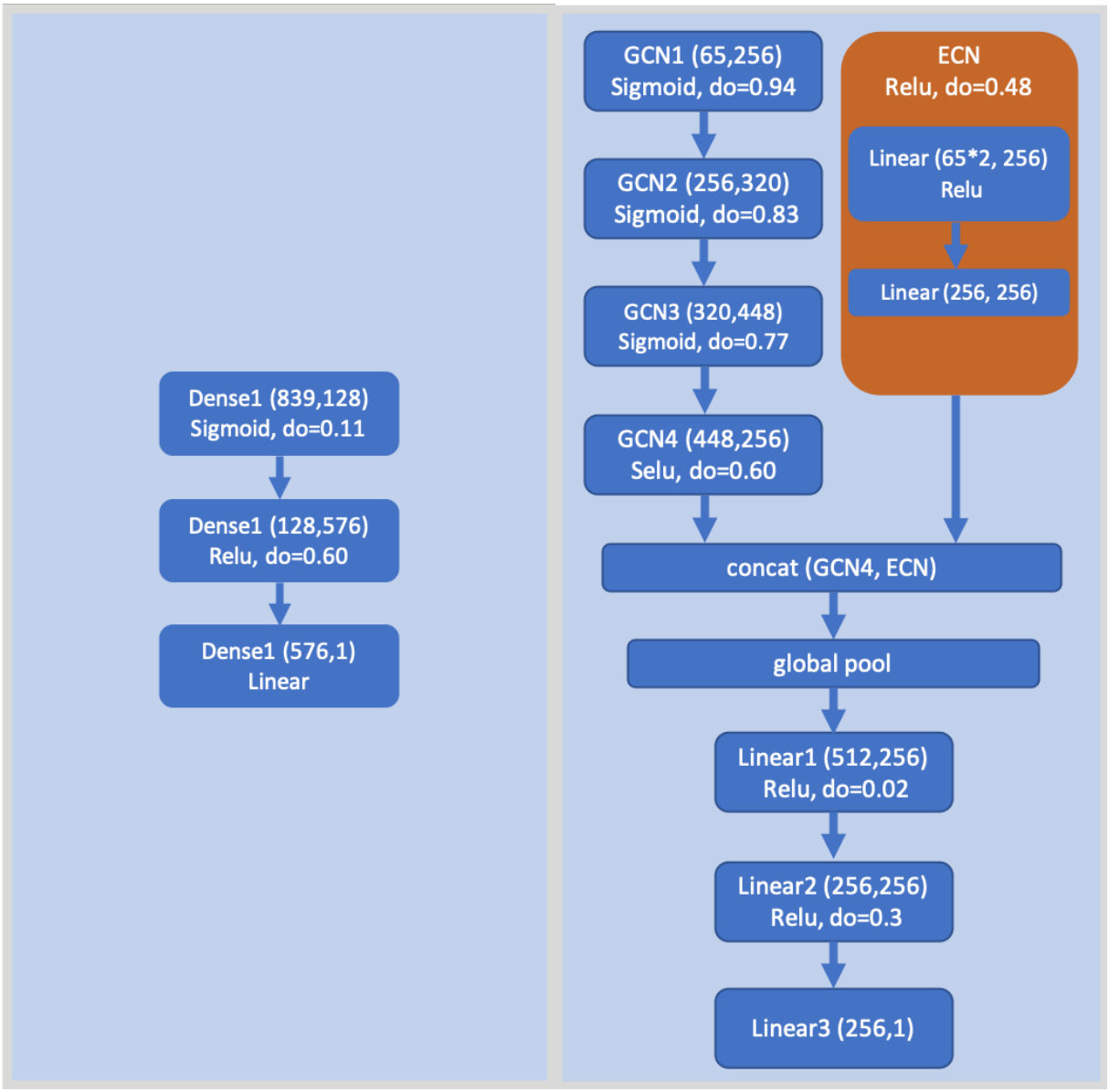
Figure S10: Binned solubility distribution

Hyper-parameter tuning

The parameters of SMILES, MDM and GNN were tuned using a combination of manual exploration and the hyperopt package.[11] For MDM and GNN models, 2000 random configurations were explored in order to find the best model. For the SMILES model, 35 configurations were considered. As SchNet is computationally expensive to train, we only changed the number of interaction layers from 6 to 12. We show the hyper-parameters and corresponding values explored and selected in Table S5. We show the final architectures of the MDM and GNN models in Figure S11.

Table S5: Hyper-parameters tuned and values explored for each model type. The selected hyper-parameters can be found in our code at https://github.com/pnnl/solubility-prediction-paper.

Model	**Hyper-parameter**	**Values**
MDM	Neurons in fully connected layers	64 to 640 by 64
	Dropout	uniform distribution 0-1
	Activation	relu, selu, sigmoid
	Learning rate	10^{-3},10^{-2}, 10^{-1}
	Optimization	adam, rmsprop, sgd
	Number of fully connected layers	2,3,4,5
GNN	Node features graph layers size	64 to 640 by 64
	Neurons in fully connected layers	32 to 320 by 32
	Dropout	uniform distribution 0-1
	Activation	relu, selu, sigmoid
	Learning rate	10^{-3},10^{-2}, 10^{-1}
	Optimization	adam, rmsprop, sgd
	Number of GCN layers	2,3,4
	Number of fully connected layers	2,3,4
3 SMI	Embedding dimension	64 to 1088 by 64
	Number of LSTM output units	64 to 576 by 64
	Neurons in fully connected layers	64 to 1088 by 64
	Dropout	uniform distribution 0-1
	Activation	relu, selu, sigmoid
	Learning rate	10^{-3},10^{-2}, 10^{-1}
	Optimization	adam, rmsprop, sgd
	Number of fully connected layers	2,3,4,5
SCH	Embedding size for atoms	64 and 128
	Number of filters	64 and 128
	Number of interactions	3 to 10
	Number of layers in MLP	1 to 4
	Aggregation mode	“sum” and “avg”



(a) MDM (b) GNN

Figure S11: Final architectures of (a) MDM (implemented using Keras), and (b) GNN (implemented using PyTorch and PyTorch geometric) models. Input and output dimensions of neural network layers are given inside parenthesis respectively. “do" stands for dropout rate.

Molecular Fragments Analysis

A comparison of average errors when molecules consisting of 1-4 fragments is tabulated in Table S6. Here, the error is defined as $|Actual - Predicted|$ solubility. In general, MDM makes better predictions for fragmented molecules than the GNN mode.

Table S6: GNN and MDM errors for fragmented molecules in the test set.

fragments	GNN error	MDM error
1	0.717437	0.694818
2	0.652585	0.569263
3	0.846883	0.744782
4	1.469782	1.538846
>1	0.718044	0.632401

Cluster Analysis

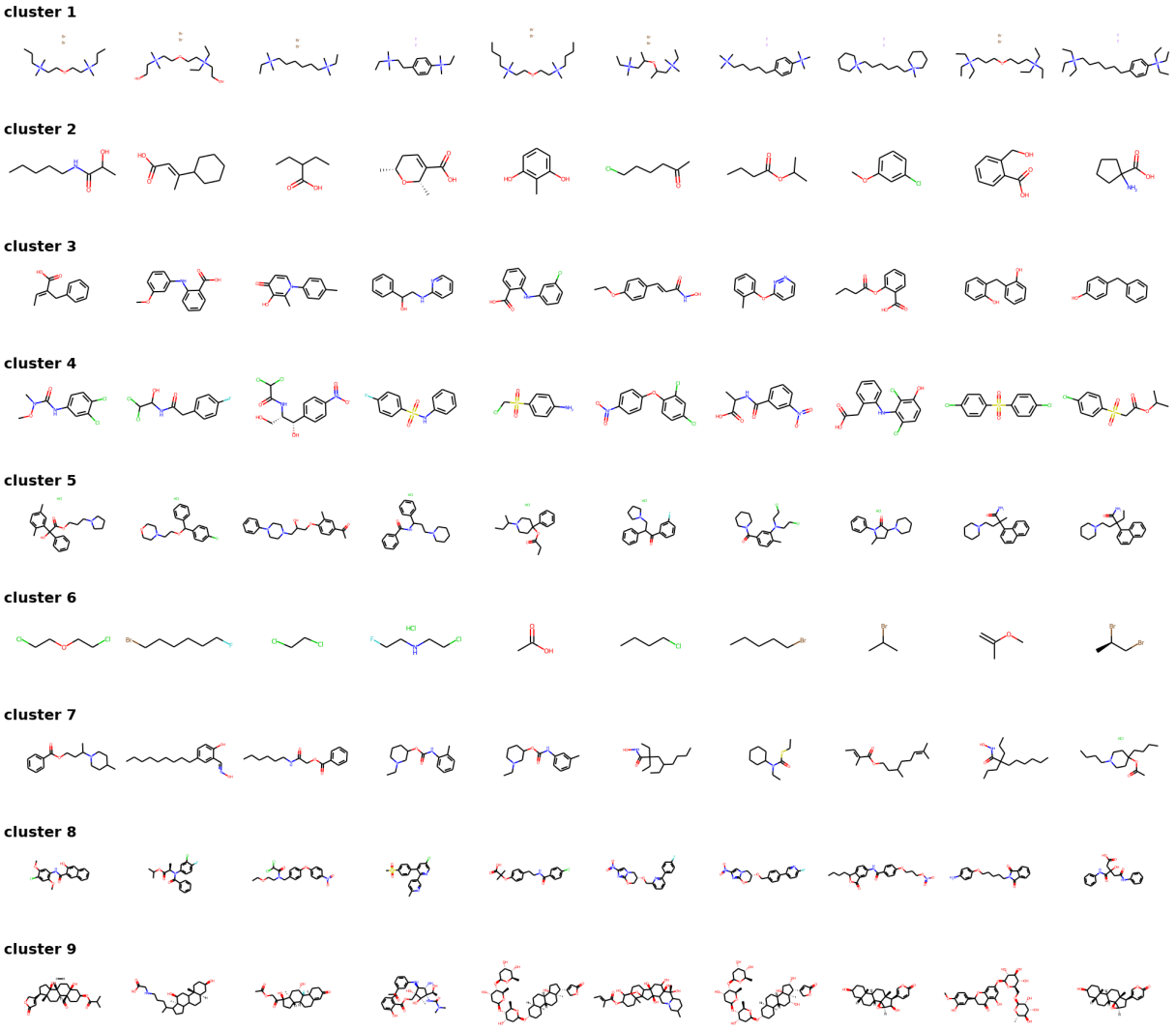


Figure S12: Ten molecules closest to the 9 cluster centers.

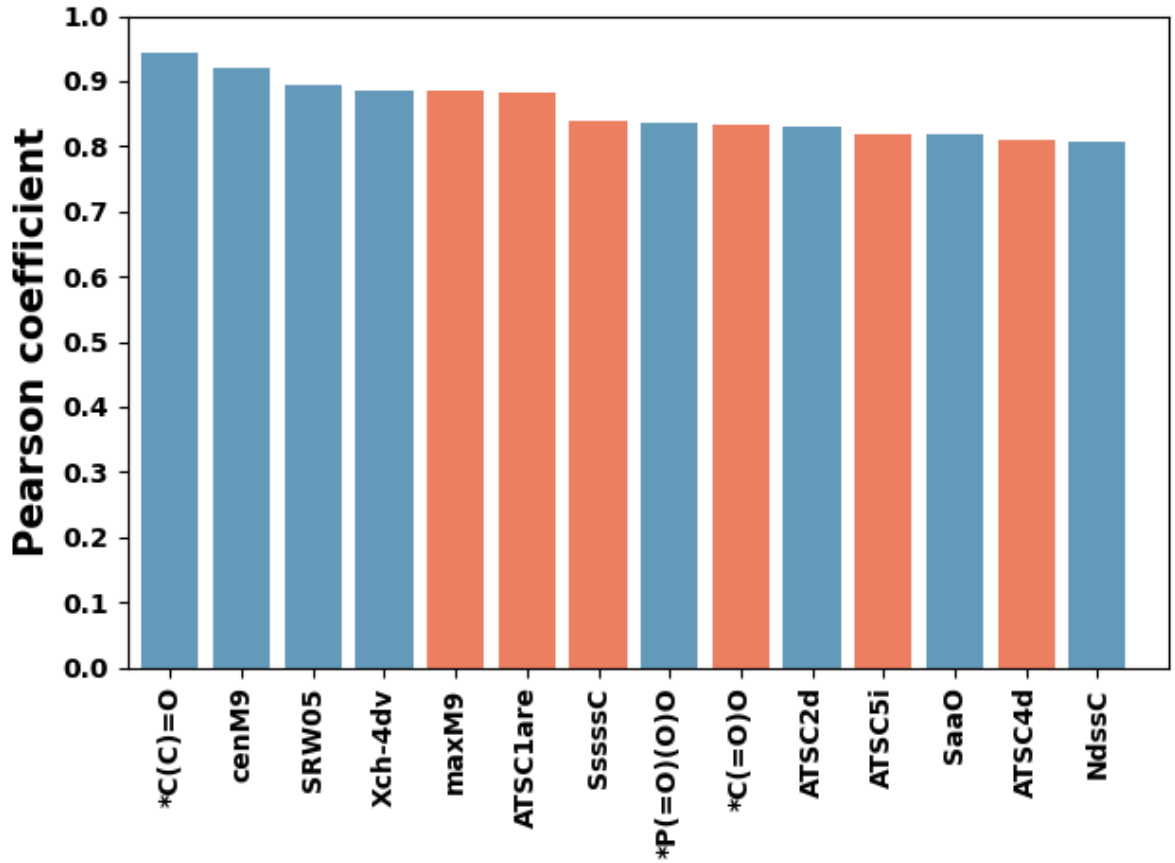


Figure S13: Highly correlated descriptors with average error corresponding to clusters.

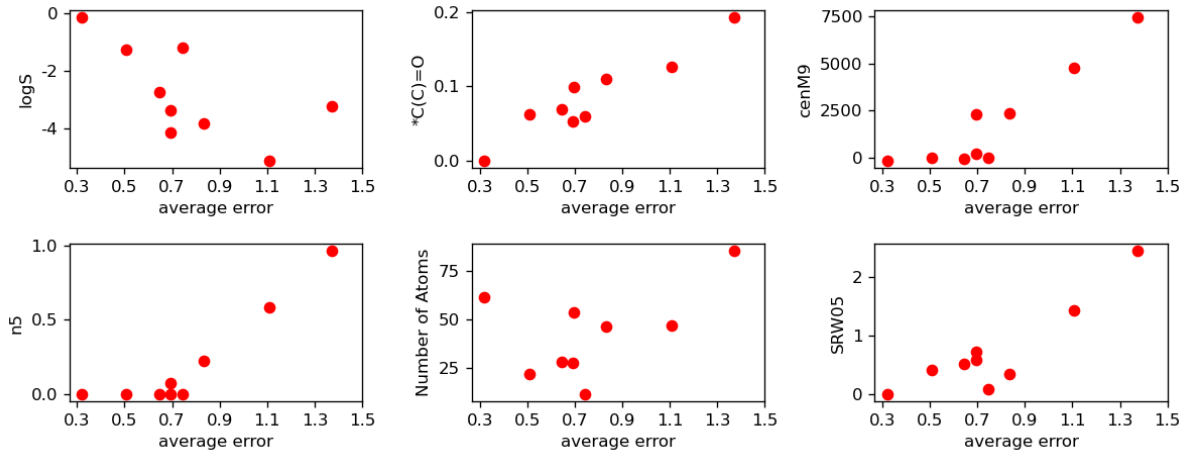


Figure S14: Scatter plot of the mean value of the descriptors versus the cluster error.

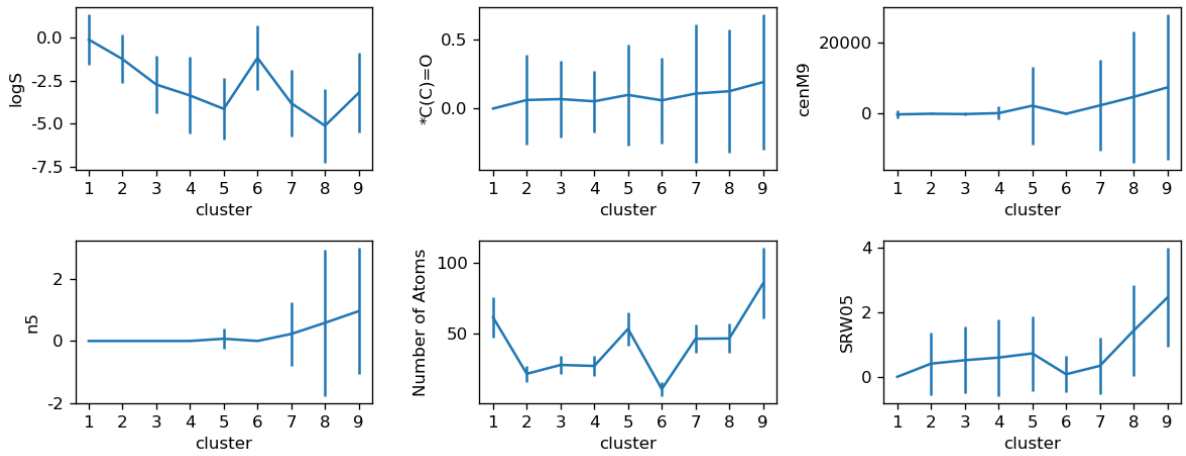


Figure S15: Scatter plot of the mean value of the descriptors versus the cluster label.

Preparation of isomer, same core and and functional group sets

Isomer set. First, molecules that have the same core structure and the same number of atoms for each constituent atom type were selected. Next, the structural similarity between each pair of molecules in the resulting groups were calculated using MACCS keys as implemented in RDKit. If the structural similarity for each pair is above a certain threshold value, the molecules in the group were registered as isomer structures.

Same core set. Core structures were found by sequentially removing molecular fragments from the molecules. The molecular fragments considered for this task and the order in which they are removed were determined based on trial and error testing. The resulting structures were then inspected by a chemist and only the structures that he confirmed as core structures were considered as core structures.

Same functional group set. We sought the help of a chemist to identify twenty one functional groups out of the most common molecular fragments attached to the molecules in our database. We then used RDKit to find molecules each functional group is associated with.

Sterimol parameters

The Sterimol parameters as originally postulated by Verloop, et al. in 1976 laid out metrics by which to gauge the sterically localized anomalies of a molecule, or substituent. The original parameters, physically measured in Angstroms, called for five directions along the parent molecule's L-axis, which is defined by the maximum tangential length running between the specified torsional bond atoms and the molecular Van der Waals volumetric shell surface.[12] This parameter was subsequently revised by the inclusion of an adjustment weight for a typical carbon-carbon bond. The revised width (B) parameters were defined as the smallest distance from the L-axis, and the maximum width as measured along coordinate system axes, making for a B1 and B5 parameter value.[13]

Other steric forms of molecular measure become equally important to consider, such as the Charton v constant, or the Taft steric parameters; however, what is most salient about Verloop's parameters is their versatility in being frequently applied to QSAR studies, as well as adapted by the inclusion of meaningful quantum information from a molecule's ab initio computational studies, such as DFT. This has been notably explored for parent-derivative compound family-based studies, but continues to lend credence for potential future use due to the heuristic consideration of varied molecular topologies in its principle: namely, the space- constrained geometry of a Van der Waals volumetric shell.[14] While the data obtained from Sterimol parameters are highly dependent on the prudent selection of a spatially centralized torsional bond within the molecule, it is safe to reason that its enhancement by way of inclusion of data from quantum data-informed computational molecular studies can prove to be quite powerful.

References

(1) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10*, 4.

(2) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data* **2019**, *6*, 143.

(3) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.

(4) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *The Journal of Organic Chemistry* **1968**, *33*, 347–350.

(5) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1000–1005.

(6) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *Journal of Chemical Information and Modeling* **2009**, *49*, 2572–2587.

(7) Priimagi, A.; Cavallo, G.; Metrangolo, P.; Resnati, G. The Halogen Bond in the Design of Functional Supramolecular Materials: Recent Advances. *Accounts of Chemical Research* **2013**, *46*, 2686–2695.

(8) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* **2016**, *abs/1609.02907*.

(9) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *CoRR* **2019**, *abs/1903.02428*.

(10) Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; Solomon, J. M. Dynamic Graph CNN for Learning on Point Clouds. *CoRR* **2018**, *abs/1801.07829*.

(11) Bergstra, J.; Yamins, D.; Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. 2013; p I–115–I–123.

(12) Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*; Ariëns, E., Ed.; Medicinal Chemistry: A Series of Monographs; Academic Press: Amsterdam, 1976; Vol. 11; pp 165–207.

(13) Verloop, A. In *Pesticide Chemistry: Human Welfare and Environment*; DOYLE, P., FUJITA, T., Eds.; Pergamon, 1983; pp 339–344.

(14) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catalysis* **2019**, *9*, 2313–2323, Publisher: American Chemical Society.