Supplementary Information (SI) for Digital Discovery.

A Machine Learning Approach for the Prediction of Aqueous Solubility of Pharmaceuticals: A Comparative Model and Dataset Analysis

Mohammad Amin Ghanavati[1], Soroush Ahmadi[1,2], Sohrab Rohani[1,*]

1 Chemical and Biochemical Engineering, Western University, London, Ontario N6A 5B9, Canada

2 Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Corresponding author: srohani@uwo.ca

S1. Sensitivity of molecular representations to stereoisomerism

In this section, we evaluate the sensitivity of three proposed molecular representations to stereoisomerism. While canonical SMILES do not distinguish between stereoisomers, isomeric SMILES capture chirality through chiral indicators within the string. Our dataset uses isomeric SMILES to preserve this information.

Chirality refers to molecules that are non-superimposable mirror images around a chiral center. In SMILES notation, chirality is denoted by @ for anticlockwise (left-handed) and @@ for clockwise (right-handed) arrangements[1]. For example, L-Alanine is represented as CC@HN, with @H indicating an anticlockwise arrangement, while D-Alanine is represented as CC@@HN, with @@H indicating a clockwise arrangement (Figure S1.a).

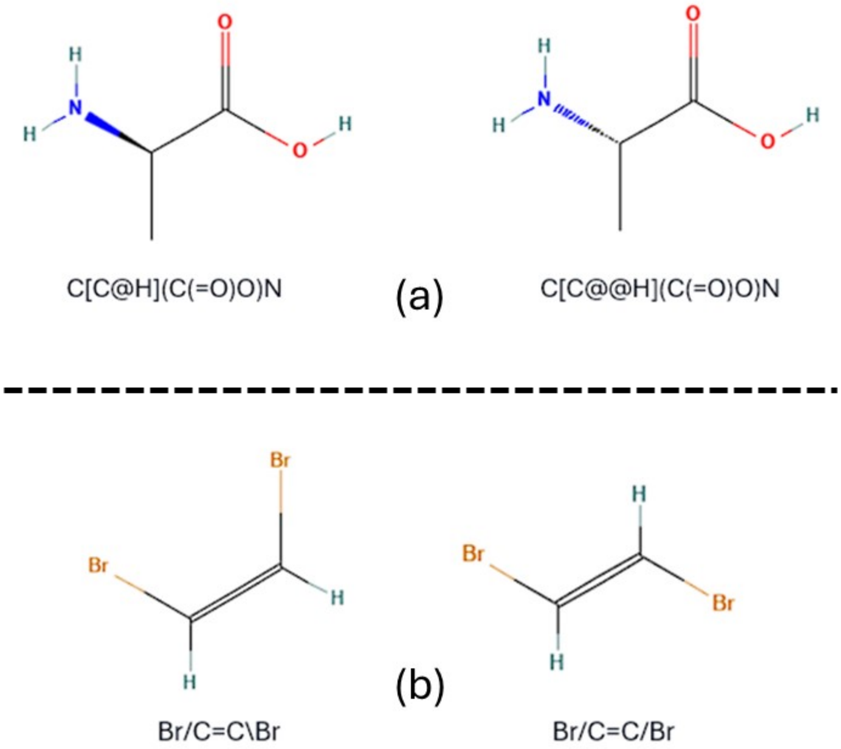

Figure S1. examples of isomeric SMILES; (a) chiral isomers and, (b) cis-trans isomers

Geometric isomers (cis-trans isomers) differ in spatial arrangement around a double bond or ring. In the cis (Z) configuration, substituents are on the same side, while in the trans (E) configuration, they are on opposite sides. The / and \ symbols in SMILES denote these configurations; / for cis and \ for trans. For instance, in the solubility dataset, Br/C=C/Br (cis (Z)) and Br/C=C\Br (trans (E)) are examples (Figure S1.b).

Among the three molecular representations, ESP maps (and their extracted features) and Mordred descriptors effectively capture stereoisomerism. In contrast, molecular graph representations, which are based solely on atom connectivity, cannot capture isomeric structures. ESP maps, obtained after geometry optimization, generate four-dimensional (x, y, z, ESP) isosurfaces unique to each isomeric structure, with features such as hydrogen bond donors/acceptors and shape characteristics sensitive to stereoisomerism. Similarly, many Mordred descriptors, based on 3D molecular characteristics, are also sensitive to stereoisomerism. Notable examples include:

- CPSA class: partial positive and negative surface area (PPSA and PNSA), fractional charged surface area (FPSA and FNSA), and surface-weighted charged surface area (WPSA and WNSA)
- Geometrical Index: geometric radius (Radius3D), geometric shape index (GeomShapeIndex), and heavy atom gravitational index (GRAV)
- Moment of Inertia: moment of inertia along the x, y, and z axes (MomentOfInertia (X, Y, Z))

The values of ESP-driven and Mordred features sensitive to stereoisomerism are detailed in Table S1 and Table S2. For additional information on Mordred descriptors, please refer to the source website.[2]

Table S1 Values of ESP map features sensitive to stereoisomerism for two example molecules.

ESP Features names	Normalized value for Br/C=C\Br	Normalized value for Br/C=C/Br
Area	0.030	0.039
Volume	0.061	0.062
Sphericity	0.412	0.331
Alpha_1	0.261	0.217
Alpha_2	0.314	0.261
Alpha_3	0.158	0.191
Alpha_4	0.169	0.204
Beta	0	0

Table S2 Values of Selected Mordred features sensitive to stereoisomerism for two example molecules.

Selected Features names	value for Br/C=C\Br	value for Br/C=C/Br
PPSA	44.63	36.38
PNSA	184.9	198.7
FPSA	0.195	0.155
FNSA	0.805	0.845
WPSA	10.24	8.55
WNSA	42.44	46.73
Radius3D	2.830	2.795
GeomShapeIndex	0.354	0.646
GRAV	1444	1165
MomentOfInertia (X)	508.74	861.1
MomentOfInertia (Y)	441.2	851.53
MomentOfInertia (Z)	67.54	9.57

S2. Performance metrics and ensemble formulas

The performance metrics used for evaluating performance of models are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) that are calculated using following equations:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{1}$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{3}$$

where N is the number of test samples in each dataset, $\hat{y}_i$ and y_i denote the predicted and true solubility values, respectively. MAE represents the average magnitude of errors, whereas RMSE is computed by taking the square root of the average of squared errors. This makes RMSE more sensitive to larger errors in predictions compared to MAE. On the other hand, R^2 represents the proportion of variance in the target variable that can be predicted from the model's predictions, ranging from 0, indicating no predictive power, to 1, representing perfect prediction.

The ensemble prediction is calculated according to the weighted summation of each model's prediction. The weights are normalized between 0 and 1 based on inverse of each model's RMSE value as follows:

$$w_j = \frac{\frac{1}{RMSE_j}}{\sum_{j=1}^{3} \frac{1}{RMSE_j}} \tag{4}$$

$$\hat{y}_{ens_j} = \sum_{i=1}^{3} w_i . \hat{y}_{ij} \tag{5}$$

In equation (4), w_j and $RMSE_j$ are weight and RMSE value of each model. In equation (4), $\hat{y}_{ens_j}$ and $\hat{y}_{ij}$ are predicted solubilities of j^{th} molecule by ensemble and i^{th} model, respectively.

S3. Top 10 tabular descriptors

Based on feature importance analysis we performed the top molecular descriptors from ESP map and Mordred are listed and described as follows:

- SLogP: Estimated lipophilicity by Mordred
- FilterltLogS: Estimated solubility by Mordred
- Beta_1: Most powerful hydrogen bond acceptor parameter (from ESP)

- RNCG: Relative negative charge
- AATS7v: AATS7v captures the averaged autocorrelation of molecular structure at a lag of 7, weighted by the Van der Waals volume of the atoms. It helps quantify how atomic environments correlate across the molecular graph, with considerations for molecular connectivity and 3D structure if applicable.
- AATS3d: captures the averaged autocorrelation of molecular structure at a lag of 3, weighted by the Sigma Electrons of the atoms. It provides insights into how the electronic environment (specifically Sigma Electrons) of atoms correlates across the molecular graph, considering molecular connectivity and optionally 3D structure.
- sph: sphericity of ESP map
- piPC2, piPC6, and piPC7: are path count descriptors that quantify the number of paths of 2, 6, and 7 bonds in length, respectively, within a molecule. These descriptors specifically focus on paths involving π-bonds, which are crucial for understanding conjugated systems. By analyzing these path counts, one can gain valuable insights into the molecule's electronic structure and the extent of its conjugation.

S4. Distribution of top features across solubility ranges

Figure S2 illustrates how second to 10th top features are distributed across various solubility ranges. The solubility ranges are defined as follows: Low logS (below the 25th percentile), Medium logS (between the 25th and 75th percentiles), and High logS (above the 75th percentile). Details on the distribution of the other top nine features are presented in Figure S2.

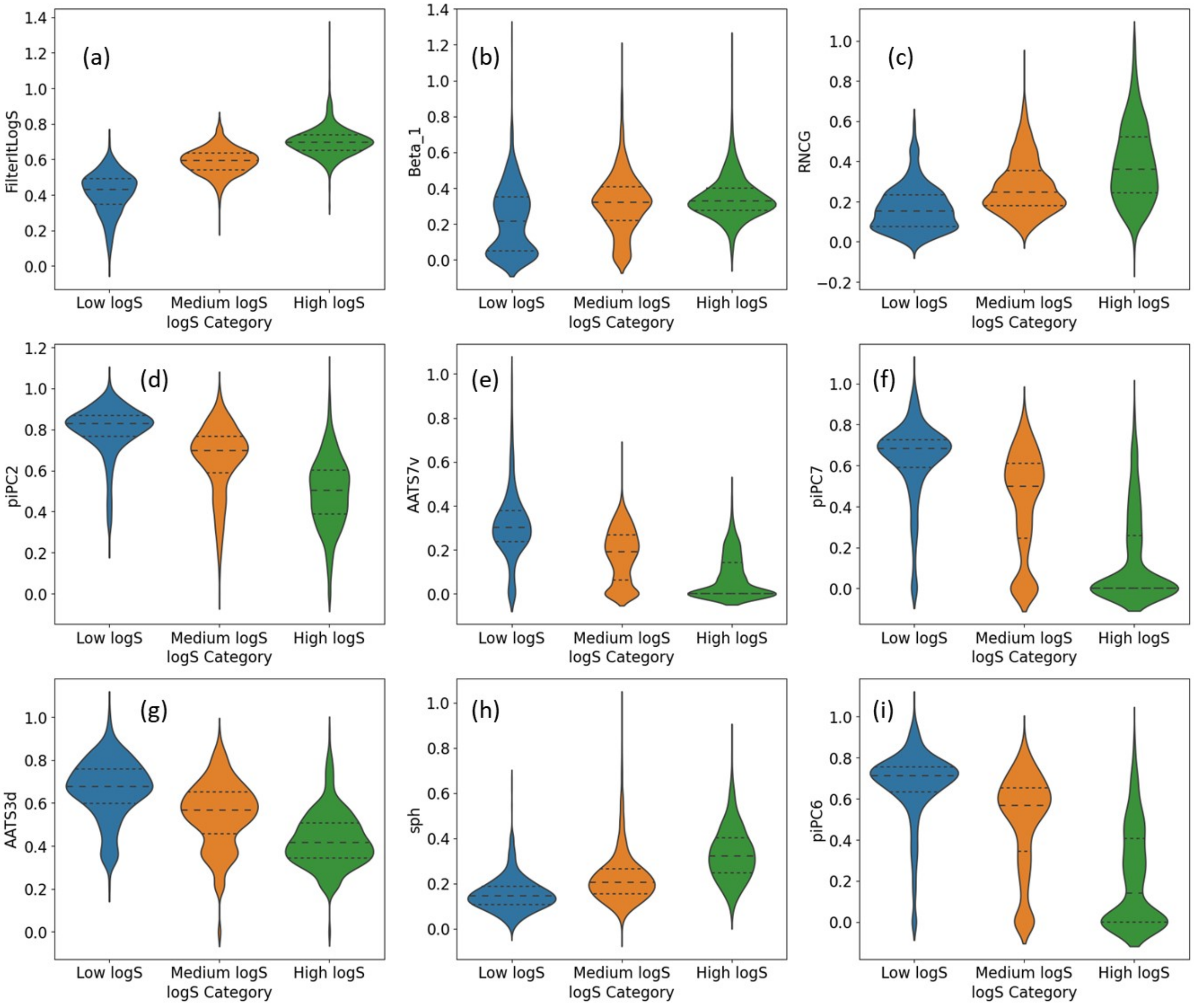

Figure S2. violin plot distribution of top remaining features across solubility ranges

S5. Comparison of models’ size, training time, and prediction time

The three models were compared in terms of the number of parameters they have which directly affect their training time. Moreover, we included the average training time of each model over four datasets in Table S3. We evaluate the size of EdgeConv and GCN models based on the number of trainable weights. XGBoost model does not have trainable weights like neural networks, as they use an ensemble of decision trees instead of backpropagation. However, model size can be adjusted by setting number of estimators (trees) and the depth of each tree, and total number of nodes which influences the model's complexity.

For training, we utilized one NVIDIA Tesla P100 GPU on the Graham compute cluster provided by Compute Canada. However, for prediction time comparison, as in practice, many users may not be able to access GPUs, we evaluated on AMD Ryzen 9 5980HX CPU processor. As we can see from the Table S3, the largest model, EdgeConv has much longer training time than the GCN and XGBoost models. However, the training and prediction time for the most complex model is still satisfactory to be used in practice.

Table S3. Comparison of models in terms of parameters, training and prediction time.

Model	**Number of Parameters**	**Average Training Time**	**Prediction Time for one sample**
EdgeConv	1,184,769	2.6 hours	1.2 seconds
GCN	53,761	128 seconds	0.0003 seconds
Feature-based (XGBoost)	estimators: 200, trees depth: 3, nodes: 2916	37 seconds	3.8e-06 seconds

S6. Comparison with small-scale machine learning methods

We evaluated small-scale machine learning models using the full set of 1,826 Mordred descriptors, without dimensionality reduction. The models assessed include K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT), all trained and tested on the same data splits across four datasets (Table S4). These results were compared with those of the three primary models based on different molecular representations—EdgeConv, GCN, and Feature-based—and their ensemble.

Among the individual small-scale models, SVM achieved the best performance. However, the ensemble model combining EdgeConv, GCN, and Feature-based approaches outperformed all individual models.

Table S4. Comprehensive evaluation of proposed models, including small-scale ones, on test splits across four datasets

Models	**ESOL**			**AQUA**			**PHYS**			**ALL-Data (EOAP)**		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
EdgeConv	0.843	0.621	0.841	0.887	0.681	0.827	0.857	0.641	0.838	0.988	0.756	0.801
GCN	0.782	0.598	0.863	0.752	0.569	0.876	0.732	0.540	0.882	0.756	0.552	0.884
Feature-based	0.602	0.449	0.919	0.595	0.449	0.922	0.578	0.439	0.926	0.677	0.495	0.906
Ensemble	**0.569**	**0.425**	**0.928**	**0.580**	**0.432**	**0.926**	**0.559**	**0.423**	**0.931**	**0.638**	**0.466**	**0.917**
KNN	0.816	0.616	0.851	0.805	0.591	0.857	0.867	0.643	0.834	0.807	0.610	0.867
SVM	0.723	0.519	0.882	0.647	0.468	0.908	0.663	0.503	0.903	0.720	0.527	0.894
DT	0.930	0.700	0.806	0.815	0.699	0.854	0.963	0.694	0.795	0.964	0.689	0.810

S7. Error distribution by functional groups

To identify challenging functional groups for model predictions, we calculated the mean error per functional group type using the Ensemble model across all unique molecules in the dataset. These errors, along with the count of each functional group, are visualized in Figure S3.

Our analysis reveals that molecules with sulfonic functional groups exhibit high error rates, primarily due to their low representation in the dataset. Phosphate-containing molecules, the second smallest group in our dataset, also show elevated errors. This is expected, as these groups are relatively rare and thus act as outliers in data-driven models. For other functional groups, error rates do not vary significantly, suggesting that no specific functional group poses a particular challenge for the predictive model beyond issues related to their limited presence.

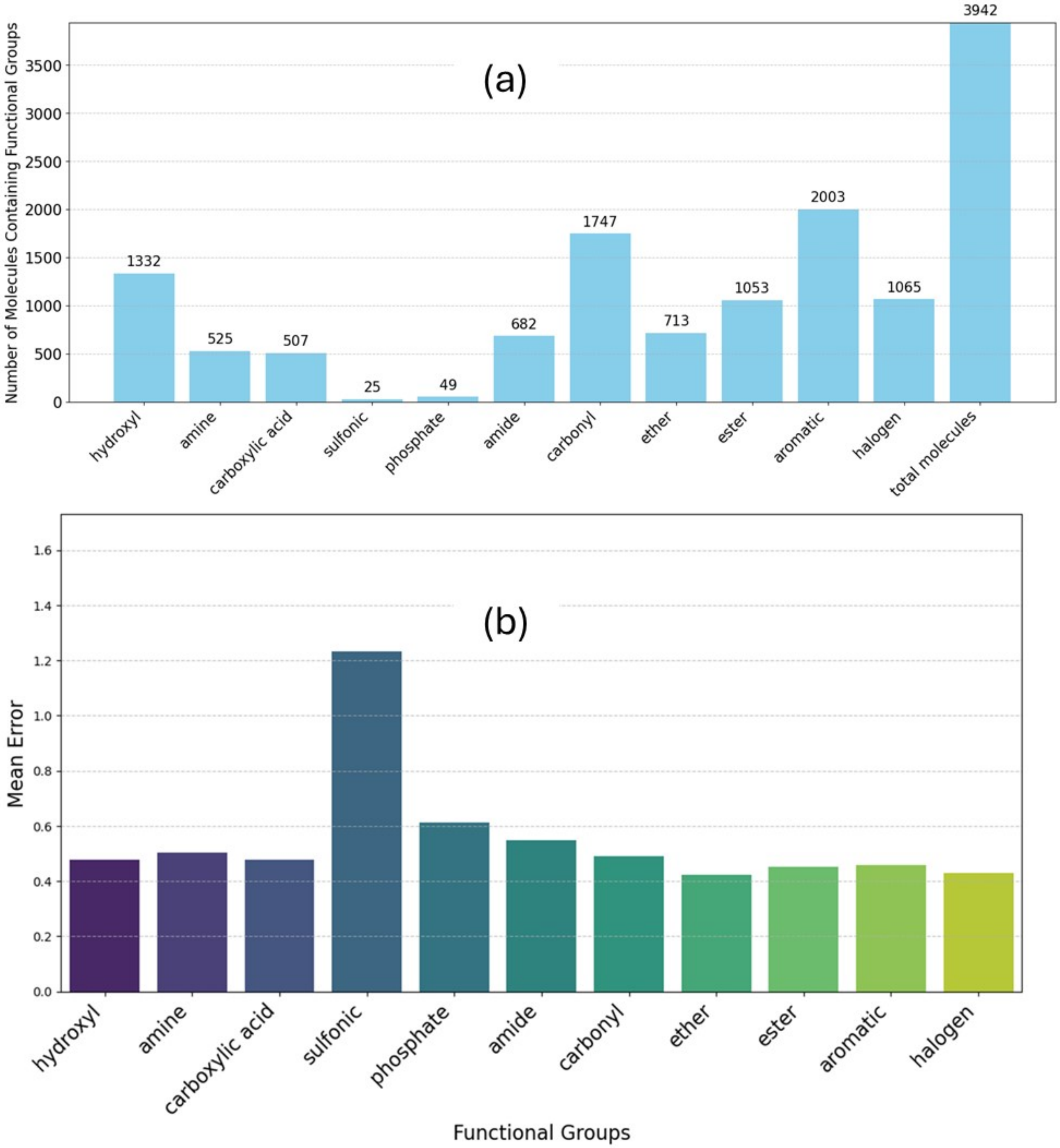

Figure S3. Error distribution across functional group types and their frequencies

S8. Additional examples of SHAP explanations

Four additional molecule pairs examples analyzed by SHAP waterfalls are visualize in Figure S4 and Figure S5.

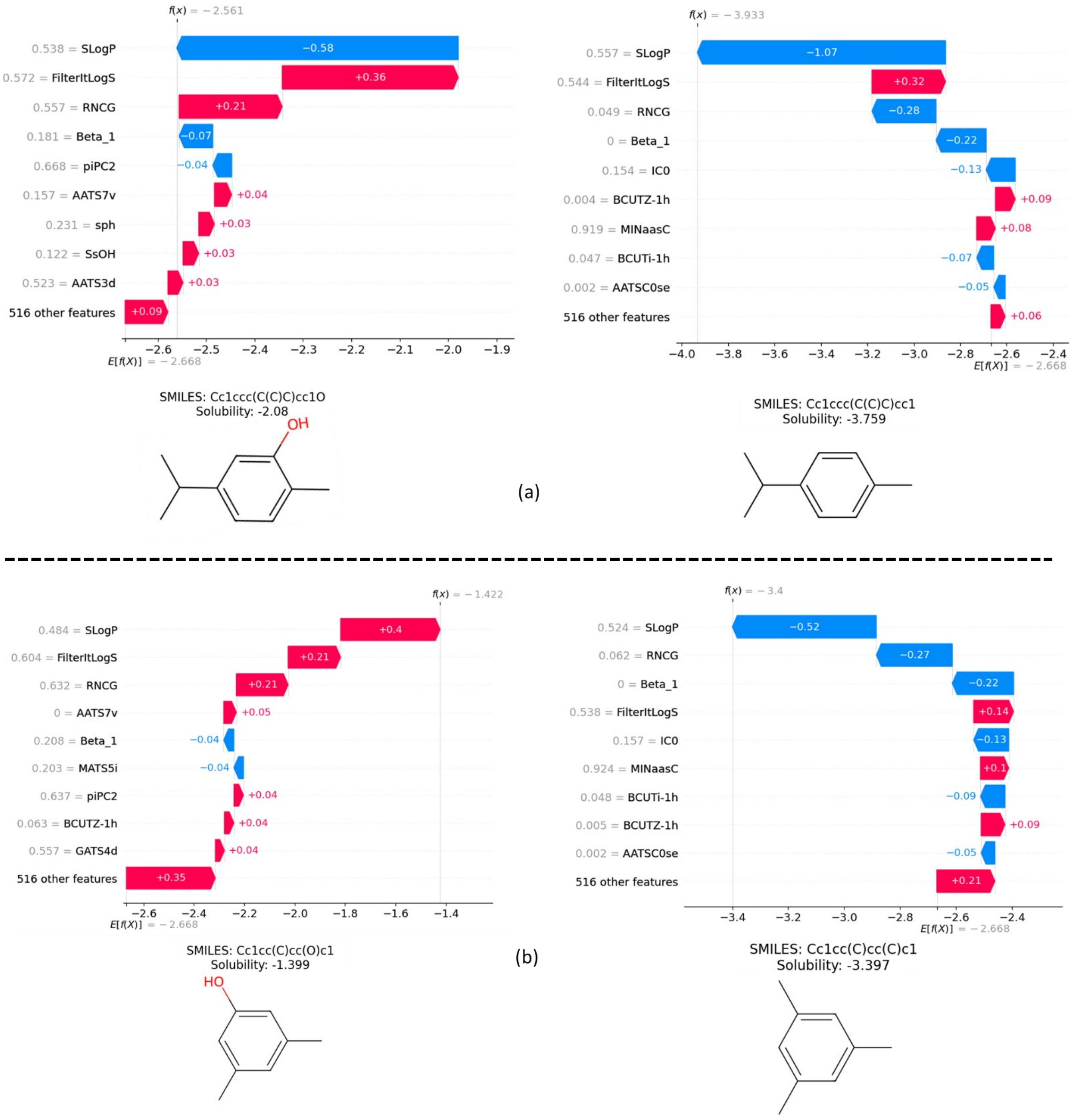

Figure S4. explanation of predictions for four pairs of molecules with feature-based model; examples a and b.

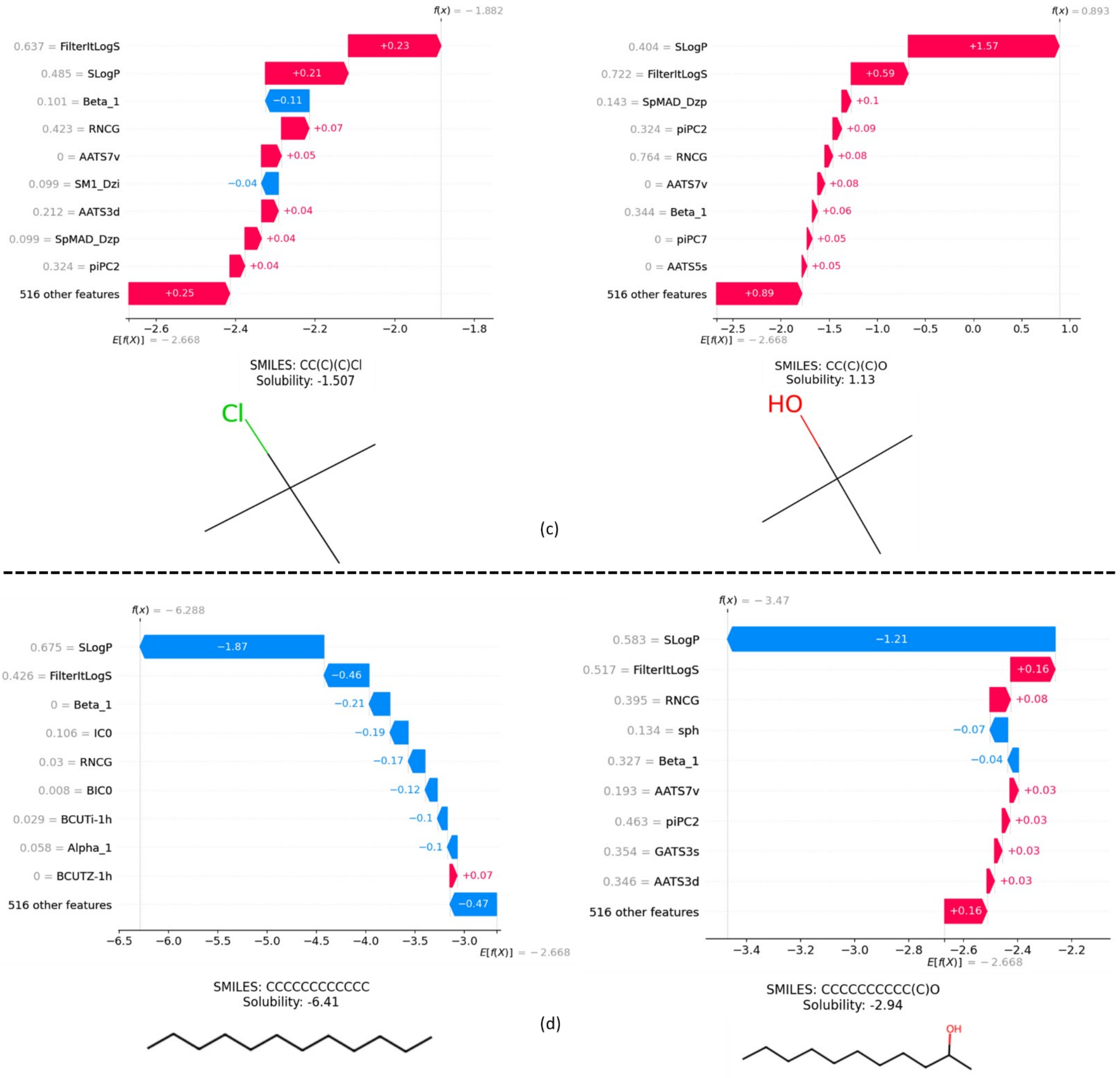

Figure S5. Continued from Figure S4: Explanation of predictions for pairs of molecules using a feature-based model; examples c and d.

References

(1) Daylight Chemical Information Systems, Inc. (n.d.). SMILES Theory (Version 1.0). Retrieved July 30, 2024, from Https://Www.Daylight.Com/Dayhtml/Doc/Theory/Theory.Smiles.Html.

(2) "Mordred Documentation." *Mordred Descriptors*, Mordred Descriptors, n.d., Https://Mordred-Descriptor.Github.Io/Documentation/Master/Descriptors.Html. Accessed 30 July 2024.