Supporting Information: Improved Environmental Chemistry Property Prediction of Molecules with Graph Machine Learning

Shang Zhu,[†] Bichlien H. Nguyen,[‡] Yingce Xia,[‡] Kali Frost,[‡] Shufang Xie,[‡] Venkatasubramanian Viswanathan,[†] and Jake A. Smith[*,‡]

†*Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*

‡*Microsoft Research, AI4Science.*

E-mail: jakesmith@microsoft.com

Mathematical Details on O-GNN

Mathematically, we can define the molecular graph G as

$$G = (\mathbf{V}, \mathbf{E}, \mathbf{R}) \tag{1}$$

where $\mathbf{V}$ is the atom set, $\mathbf{V} = \{v_1, v_2, ..., v_{|\mathbf{V}|}\}$, $\mathbf{E}$ is the bond set, $\mathbf{E} = \{e_{ij} | i, j \in |\mathbf{V}|\}$, and $\mathbf{R}$ is the ring set, $\mathbf{R} = \{r_1, r_2, ..., r_{|\mathbf{R}|}\}$. r_i is defined as a simple ring (i.e. including only one ring). Atom, bond, and ring features are further specified by $h^{(0)}_{v_i}$ (atom type, chirality, degree number, etc.), $h^{(0)}_{e_{ij}}$ (bond type, stereochemistry, conjugated type) and $h^{(0)}_{r_i}$ (a concatenation of atom and bond features that are involved in the rings). The superscript $^{(0)}$ represents that $h^{(0)}$ is the 0^{th} layer feature, i.e. the input feature. In the iterative message passing step, we can obtain the bond features, atom features, ring features and additionally a compound feature $U^{(l)}$ at l^{th} layer by following the below update steps sequentially:

$$h_{e_{ij}}^{(l)} = h_{e_{ij}}^{(l-1)} + \texttt{Aggregate}(h_{e_{ij}}^{(l-1)}, h_{v_i}^{(l-1)}, h_{v_j}^{(l-1)}, \{h_{r_{i'}}^{(l-1)}\}_{i' \in R(e_{ij})}, U^{(l-1)}) \tag{2}$$

$$h_{v_i}^{(l)} = h_{v_i}^{(l-1)} + \texttt{Aggregate}(h_{v_i}^{(l-1)}, \{h_{e_{ij}}^{(l)}\}_{j \in N(v_i)}, \{h_{r_{i'}}^{(l-1)}\}_{i' \in R(v_i)}, U^{(l-1)}) \tag{3}$$

$$h_{r_i}^{(l)} = h_{r_i}^{(l-1)} + \texttt{Aggregate}(h_{r_i}^{(l-1)}, \{h_{v_i}^{(l)}\}_{i \in V(r_i)}, \{h_{e_{ij}}^{(l)}\}_{j \in E(r_i)}, U^{(l-1)}) \tag{4}$$

$$U^{(l)} = U^{(l-1)} + \texttt{Aggregate}(U^{(l-1)}, \{h_{v_i}^{(l)}\}_{i \in |\mathbf{V}|}, \{h_{e_{ij}}^{(l)}\}_{i,j \in |\mathbf{V}|}, \{h_{r_{i'}}^{(l)}\}_{i' \in |\mathbf{R}|}) \tag{5}$$

where $R(e_{ij}), R(v_i)$ denotes the ring sets that involve bond e_{ij} and atom v_i, respectively, while $N(v_i)$ is the neighbor atom set that connects v_i. $V(r_i)$, $E(r_i)$ represents all atoms and bonds that appear in the ring r_i. The aggregate function, $\texttt{Aggregate}(\cdot)$, is designed to convolve the information over different objects and then add them to the features from the previous $(l-1)^{th}$ layer. After L message-passing layers, we compute the mean values of transformed atom features as the graph-level molecular feature to continue with.

$$h_G^L = \frac{\Sigma_{i \in |\mathbf{V}|} h_{v_i}^{(l)}}{|\mathbf{V}|} \tag{6}$$

The environmental-related molecular properties can then be obtained by transforming this graph-level feature with a multi-layer-perceptron network, $\texttt{MLP}(\cdot)$.

$$f(h_G^L) = \texttt{MLP}(h_G^L) \tag{7}$$

Algorithm Implementation and Model Selection

For the conventional-feature-based models, we created the ECFP and MACCS features with RDKit[1] built-in functions (`rdkit.Chem.AllChem.GetMorganFingerprintAsBitVect()`, `rdkit.Chem.MACCSkeys.GenMACCSKeys()`), while the Mordred GitHub repository[2] (`https://github.com/mordred-descriptor/mordred`) was used for generating Mordred descriptors. ECFP features were generated with the radius of 2 and the number of bits of 2048 (i.e. 2048-dimension feature vector), and MACCS features were of 167 dimensions. For Mordred features, only 2D descriptors were calculated with up to 1613 dimensions. Machine learning algorithms were imported from `scikit-learn` package.[3] We also performed the standard scaling (remove means and normalize it to unit variance) with `sklearn.preprocessing.StandardScaler` as a model variant, which may potentially improve the regression accuracy. In terms of algorithms, the radial basis function kernel was used for support vector regression ($\gamma = 1/N_{features}$, $N_{features}$ is the number of feature dimensions.). The number of tree-based estimators and maximum depth is set as 100 and 30/3, respectively, for both random-forest and gradient-boosting regressors. Lastly, a two-layer neural network was implemented with the initial learning rate of 0.001 in the `scikit-learn` package. For all of these algorithms, unless specified, we used 5-fold cross-validation (training:testing in 8:2 ratio)and reported their root-mean-square-error on the testing set ($RMSE_{test}$) with the mean and standard deviation values.

For `NeuralFP` and `O-GNN`, we controlled the same 5-fold splits as conventional-feature-based methods to get a fair comparison. In each of the 5 splits, we trained an ensemble of 5 models by feeding the algorithm with different subsets of the training data points. The average values of this ensemble of models were used for predicting the testing data points.

NeuralFP was implemented in DeepChem.[4] We optimized NeuralFP's hyperparameters of number of message passing layers and corresponding hidden dimensions ([64,64], [128,128], [64,64,64]), dimension of the feed forward neural network layer (64, 128), as well as the dropout ratio (0, 0.2). O-GNN was implemented with PyTorch and PyTorch-Geometric. More implementation details can be found in our repository (https://github.com/shangzhu-cmu/envchemGNN.git).

Model Selection for Feature-based Models

Here are the model selections for each task, by combining chemical features with best-performing machine learning models.

Table S 1: Model Performances for Environmental Engineering Tasks[a]

Task	ESOL (1128)	BCF (1034)	Clint (4422)	O3-react (759)	SO4-react (557)
Feature	raw Modred	raw Modred	raw Modred	scaled MACCS	scaled Mordred
Selected Algorithm	Gradient Boosting	Gradient Boosting	Random Forest	Neural Networks	Support Vector Machines
$RMSE_{test}$	0.61 (0.04)	0.67 (0.05)	0.86 (0.05)	2.05	0.60

Supporting Figures

The collected environmental datasets are mapped in Figure S 1 by conducting a principal component analysis (PCA) on their molecular fingerprints.[5] The *Clint* dataset covers the broadest chemical space, compared with others that are similarly clustered in the PCA plot. The data distribution after proper transformations is visualized in Figure S1.

Figure S 2 displays the data distribution histogram after logarithm transformation, and they are the training data for benchmarked ML models.All datasets undergo logarithm transformations to better represent the data that span over orders of magnitudes. Note that the Clint dataset includes raw values of 0, and, in order to avoid numerical errors, the whole dataset has been shifted up by a negligible amount (0.0001) before the logarithm transformation.

Figure S 3 provides a few example molecules with many rings in the training datasets, with the number of rings up to 9. These indicate a significant role that ring structures may play in determining molecular properties.

Figure S 4 and S 5 show the detailed analysis for the *ESOL* task and the *BCF* task, respectively. Their pairity plots and residual loss analysis are both similar to the *Clint* task. In terms of the PCA plots, for the *ESOL* task, `O-GNN` features in Figure S 4d better distinguish the chemicals with low and high solubility labels, compared with Mordred features in Figure S 4c. The *BCF* task PCA result is slightly more complicated, since both Figure S 5c and 5d show difficulty separating the high and low bioconcentration factor values. This may be attributed to the fact that the features were extracted from one single model from the cross-validation ensemble, so the result is potentially stochastic. Further, we also observed a higher standard deviation of the *BCF* task in Table 2 of main text, than the other two tasks evaluated. Overall, after averaging from the model ensembles, `O-GNN` model is still preferred than the best feature-based machine learning model.

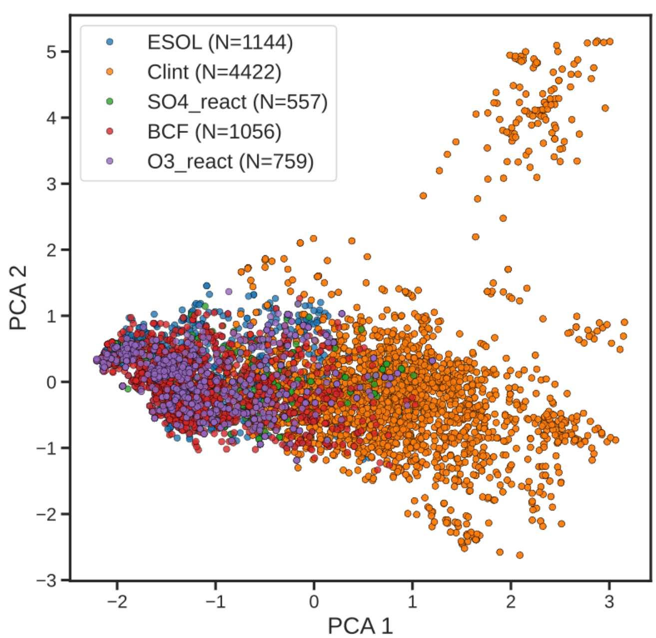

Figure S 1: Chemial Space Coverage of Curated Datasets

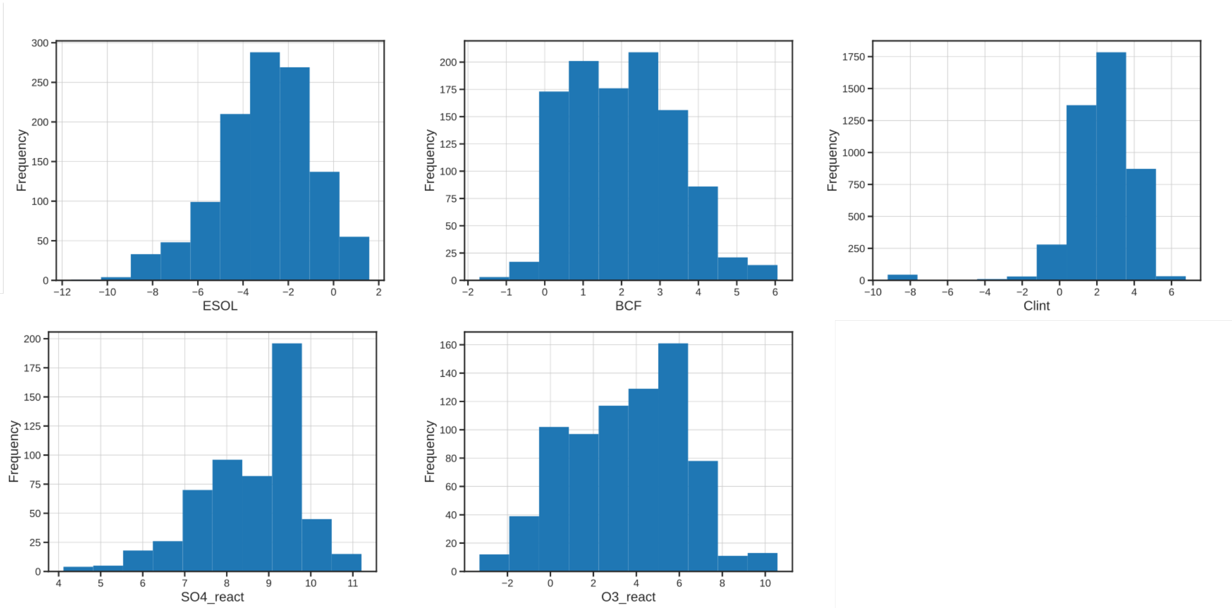

Figure S 2: Data Distribution after Logarithm Transformation

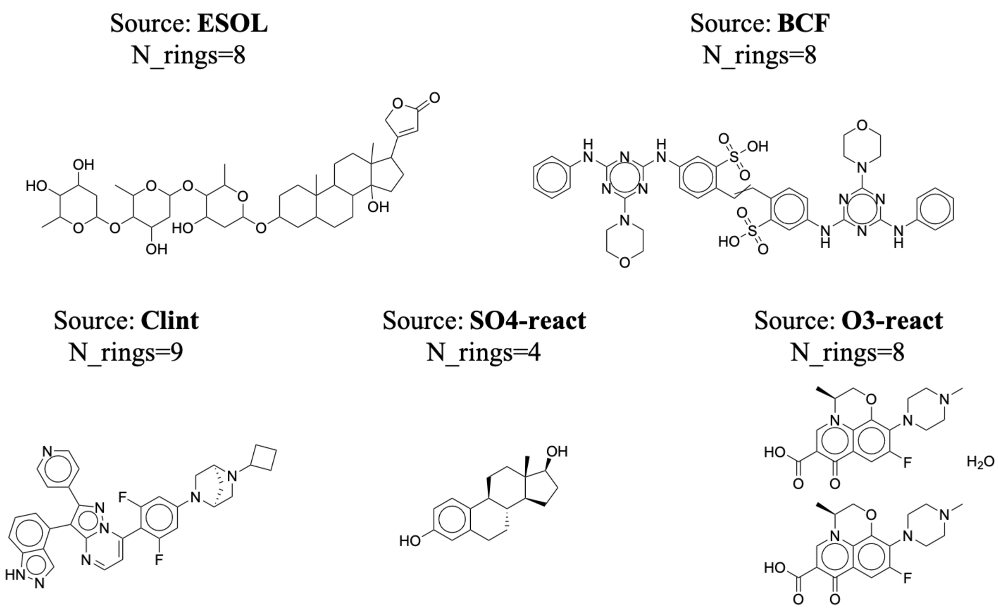

Figure S 3: Example Molecules with Large Numbers of Rings in Each Dataset

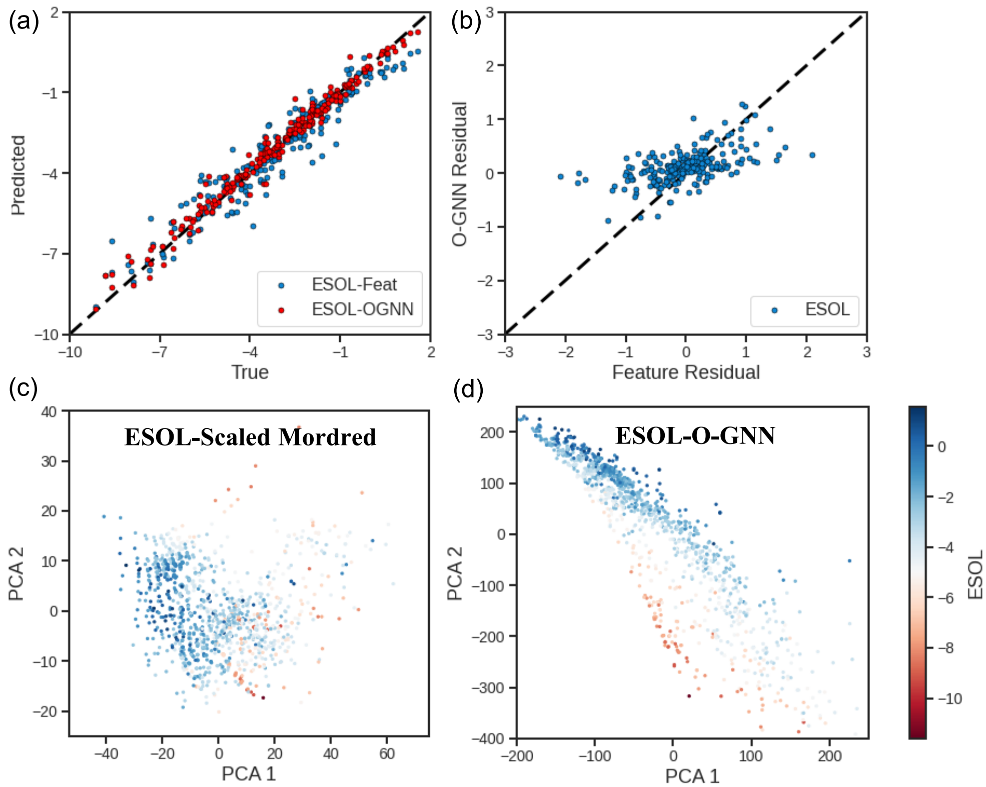

Figure S 4: Detailed Analysis of the *ESOL* task (a) Parity Plot. The black line represents complete agreement of the predicted and true values. (b) Prediction Residual Plot (predicted values minus true values). X-axis is the residual values of feature-based models while Y-axis is for `O-GNN`. (c-d) PCA Plots for (c) Scaled Mordred Features and (d) `O-GNN`-extracted Features. Each dot is color-coded by their clearance values. The scales of principal components in (c-d) depend on the raw feature scales before PCA, so the axes of these two plots are in different ranges.

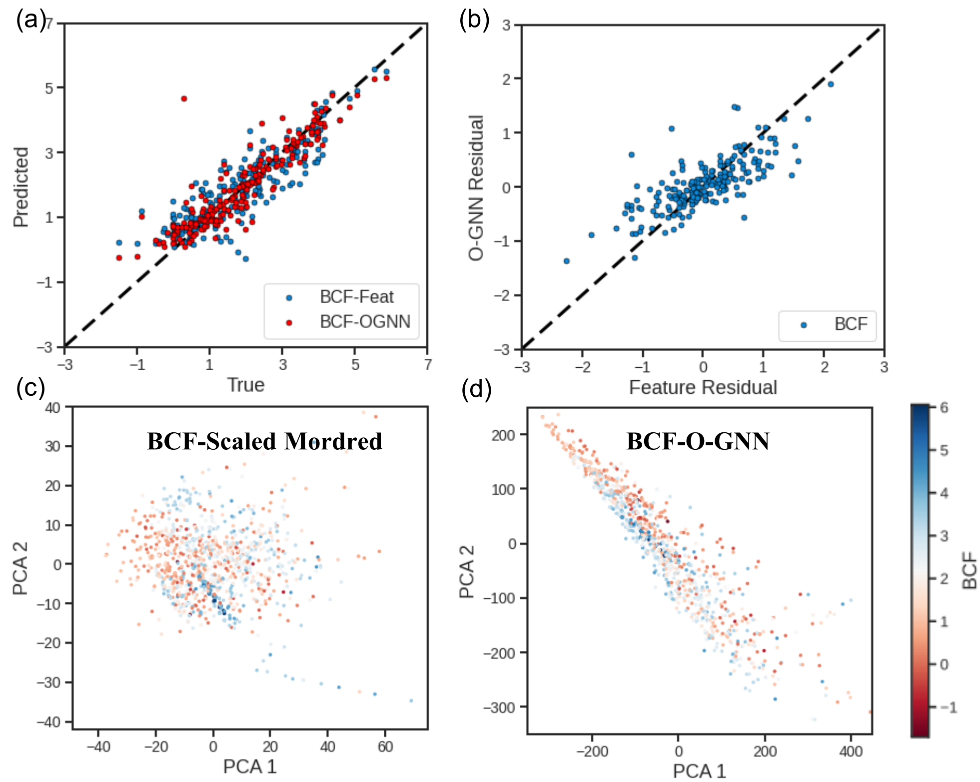

Figure S 5: Detailed Analysis of the BCF task (a) Parity Plot. The black line represents complete agreement of the predicted and true values. (b) Prediction Residual Plot (predicted values minus true values). X-axis is the residual values of feature-based models while Y-axis is for `O-GNN`. (c-d) PCA Plots for (c) Scaled Mordred Features and (d) `O-GNN`-extracted Features. Each dot is color-coded by their clearance values. The scales of principal components in (c-d) depend on the raw feature scales before PCA, so the axes of these two plots are in different ranges.

References

(1) RDKit: Open-source cheminformatics. 2020; `http://www.rdkit.org`.

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

(3) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(4) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`.

(5) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, PMID: 20426451.