# MoGAT: Multi-order Graph Attention Network for Water Solubility Prediction and Interpretation

## SUPPLEMENTARY INFORMATION

## Additional Experimental Results

Here, we report some additional experimental results to verify the effectiveness of MoGAT.

### Other Molecular Properties Prediction

We first extended experiments with some benchmark datasets to validate the predictive performance of MoGAT for other molecular properties. The datasets included four regression tasks and seven classification tasks. The regression datasets consist of *Solubility*, *FreeSolv*, *Lipop*[1], and *QM9*[2]. The target of Solubility is water solubility, and those of FreeSolv and Lipop are solvation-free energy (kcal/mol) and lipophilicity (measured octanol/water distribution coefficient, logD), respectively. QM9 is a dataset for predicting 12 chemical properties of stable organic compounds consisting of hydrogen, carbon, oxygen, nitrogen, and fluorine. In specific, the targets of QM9 include dipole moment (DM, unit: Debye), isotropic polarization (IP, unit: Bohr$^3$), highest occupied molecular orbital (HOMO, unit: Hartree), lowest unoccupied molecular orbital (LUMO, unit: Hartree), band gap energy (BGE, unit: Hartree), electronic spatial range (ESR, unit: Bohr$^2$), zero vibration energy (ZVE, unit: Hartree), internal energies at 0K and 298.15K (IE, unit: Hartree), enthalpy (unit: Hartree), Gibbs free energy (GFE, unit: Hartree), and heat capacity (HC, unit: cal/mol×K). The SMILES codes, which indicated molecular formulas and chemical structures as character strings, were given 130,000 instances in QM9 dataset, with the 12 target values for each data. In regression tasks, for comparison with AttentiveFP, we separated the QM9 dataset from the five regression datasets. For classification tasks, we used datasets related to bioactivity, physiology, and toxicity, including *MUV*, *HIV*, *BACE*, *Tox21*, *Toxcast*, *SIDER*, and *ClinTox*[1]. These tasks are binary classification, classifying whether a virus or toxic reaction exists.

**Table S1.** MAE on QM9 dataset for 12 chemical properties. The mean and standard deviation of repeated results are presented. For each target, the best performance is highlighted in boldface.

| Type (unit) | GCN | Weave | MPNN | AttentiveFP | MoGAT |
|---|---|---|---|---|---|
| DM (Debye) | 0.509 ± 8.0e-3 | 0.603 ± 3.9e-2 | 0.425 ± 4.7e-2 | 0.440 ± 1.1e-2 | **0.417 ± 7.0e-3** |
| IP (Bohr$^3$) | 0.746 ± 7.5e-2 | 1.198 ± 1.3e-1 | 0.675 ± 1.7e-1 | 0.468 ± 2.5e-2 | **0.455 ± 2.6e-2** |
| HOMO (Hartree) | 0.005 ± 2.4e-4 | 0.006 ± 2.8e-4 | 0.004 ± 8.5e-4 | 0.004 ± 3.0e-5 | **0.003 ± 6.0e-5** |
| LUMO (Hartree) | 0.005 ± 3.0e-5 | 0.007 ± 5.9e-4 | 0.005 ± 1.0e-3 | 0.004 ± 1.2e-4 | **0.004 ± 1.0e-4** |
| BGE (Hartree) | 0.007 ± 5.5e-4 | 0.008 ± 2.5e-4 | 0.006 ± 1.4e-3 | 0.005 ± 9.0e-5 | 0.005 ± 7.0e-5 |
| ESR (Bohr$^2$) | 46.511 ± 2.4e-0 | 50.684 ± 1.8e-0 | 29.517 ± 1.3e-0 | 28.363 ± 1.5e-0 | **26.378 ± 1.1e-0** |
| ZVE (Hartree) | 0.002 ± 6.2e-4 | 0.003 ± 1.3e-3 | 0.002 ± 4.9e-4 | **0.001 ± 1.0e-5** | 0.001 ± 1.0e-4 |
| IE 0K (Hartree) | 2.226 ± 2.4e-1 | 2.126 ± 3.7e-1 | 1.672 ± 4.9e-1 | 0.874 ± 2.5e-2 | **0.794 ± 1.4e-1** |
| IE 298.15K (Hartree) | 2.231 ± 2.5e-1 | 2.124 ± 3.8e-1 | 1.665 ± 4.9e-1 | 0.871 ± 2.2e-2 | **0.794 ± 1.4e-1** |
| Enthalpy (Hartree) | 2.230 ± 2.5e-1 | 2.118 ± 3.8e-1 | 1.658 ± 4.8e-1 | 0.871 ± 2.2e-2 | **0.794 ± 1.4e-1** |
| GFE (Hartree) | 2.230 ± 2.4e-1 | 2.121 ± 3.8e-1 | 1.663 ± 4.8e-1 | 0.871 ± 2.2e-2 | **0.794 ± 1.4e-1** |
| HC (cal/mol·K) | 0.338 ± 8.0e-3 | 0.529 ± 6.5e-2 | 0.310 ± 8.6e-2 | **0.235 ± 1.8e-2** | 0.240 ± 8.0e-3 |

Tables S1 and S2 list mean absolute error (MAE) of the QM9 and RMSE of the rest regression tasks, respectively. In both the QM9 and the rest datasets, MoGAT showed a better prediction than the other existing methods.

Table S3 indicates AUROC scores for the classification tasks. MoGAT showed superior performance compared to the other baseline methods in five out of seven classification tasks.

### Water Solubility Prediction

Next, for various molecules, we compared AttentiveFP and MoGAT in terms of the importance scores for each chemical component in a molecule and predicted water solubility values. In Figures S1-S5, the component symbol of carbon and hydrogen connected to carbon are omitted; the names and chemical formulas of the molecules are indicated at the top, and the target values are presented bottom; the results of MoGAT is shown on the left, and that of AttentiveFP on the right.

**Table S2.** RMSE on the rest regression tasks except for QM9. For each dataset, the best performance is highlighted in boldface.

| Type | Solubility (unit: logS) | Freesolv (unit: kcal/mol) | Lipop (unit: logD) |
|---|---|---|---|
| GCN | 0.723 | 1.130 | 0.990 |
| Weave | 0.460 | 1.304 | 1.125 |
| MPNN | 0.350 | 1.113 | 0.913 |
| AttentiveFP | 0.286 | 0.920 | 0.617 |
| MoGAT | **0.281** | **0.835** | **0.599** |

**Table S3.** AUROC on classification tasks. This score indicates that the closer to 1, the better the prediction performance. The mean and standard deviation of repeated results are presented. For each dataset, the best performance is highlighted in boldface.

| Type | HIV | BACE | MUV | Toxcast | Tox21 | SIDER | ClinTox |
|---|---|---|---|---|---|---|---|
| GCN | 0.788 ± 0.011 | 0.863 ± 0.011 | 0.709 ± 0.062 | 0.710 ± 0.003 | 0.819 ± 0.004 | 0.592 ± 0.015 | 0.841 ± 0.020 |
| Weave | 0.501 ± 0.002 | 0.755 ± 0.012 | 0.500 ± 0.002 | 0.723 ± 0.004 | 0.793 ± 0.002 | 0.555 ± 0.019 | 0.852 ± 0.031 |
| MPNN | 0.742 ± 0.028 | 0.869 ± 0.009 | 0.485 ± 0.021 | 0.733 ± 0.004 | 0.816 ± 0.006 | 0.585 ± 0.029 | 0.848 ± 0.026 |
| Attentive FP | 0.795 ± 0.026 | **0.874** ± **0.012** | **0.806** ± **0.027** | 0.835 ± 0.006 | 0.844 ± 0.009 | 0.620 ± 0.017 | 0.943 ± 0.007 |
| MoGAT | **0.821** ± **0.003** | 0.873 ± 0.015 | 0.804 ± 0.016 | **0.851** ± **0.006** | **0.846** ± **0.012** | **0.641** ± **0.015** | **0.964** ± **0.012** |

The predictive performances for some molecules of the proposed MoGAT achieved slightly lower than those of AttentiveFP. However, MoGAT provided more reasonable importance scores than AttentiveFP.

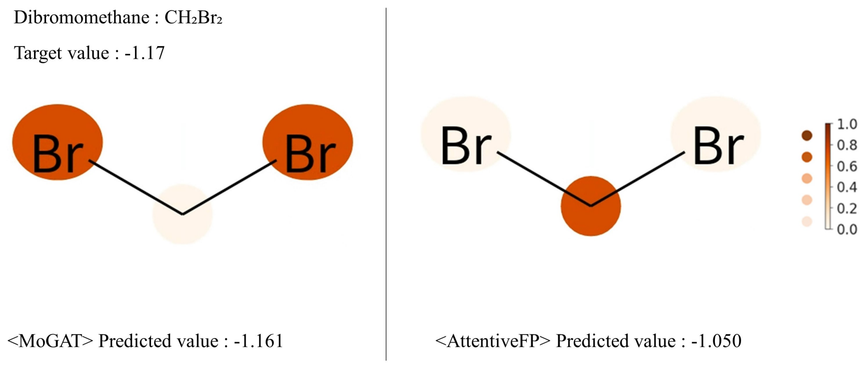
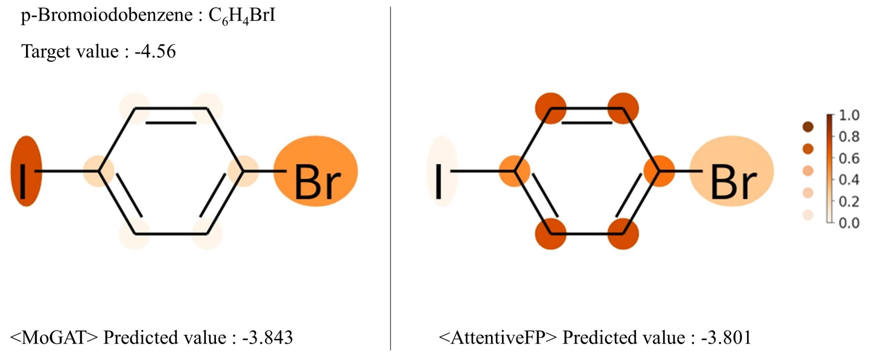

Dibromomethane : CH2Br2
Target value : -1.17
Br
Br
Br
Br
1.0
0.8
0.6
0.4
0.2
0.0
<MoGAT> Predicted value : -1.161
<AttentiveFP> Predicted value : -1.050

**Figure S1.** The predicted water solubility (unit: logS) and importance of each atom for dibromomethane molecule. The attention scores are presented by the color coding in the right panel.

## Replacement of Some Components in Molecules

In this section, for various molecules, we confirmed how the predicted values and attention scores changed when some chemical components of the molecules were replaced with others. For example, O, $OH_x$, and $NH_x$ are replaced with $CH_x$, or $CH_x$ was substituted with O or $OH_x$. In Figures S6-S13, the component symbol of carbon and hydrogen connected to carbon were omitted. In addition, the names and chemical formulas of the molecules are indicated at the top, and the target values are presented bottom. Finally, the substituted atoms and chemical formulas before and after transformation were shown for each molecule.

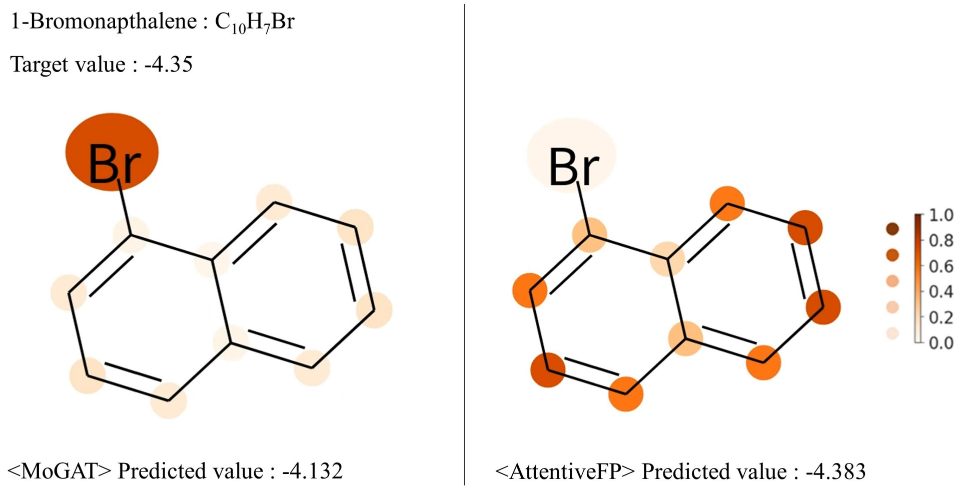

p-Bromoiodobenzene : $C_6H_4BrI$

Target value : -4.56

<MoGAT> Predicted value : -3.843

<AttentiveFP> Predicted value : -3.801

**Figure S2.** The predicted water solubility (unit: logS) and importance of each atom for p-bromoiodobenzene molecule. The attention scores are presented by the color coding in the right panel.

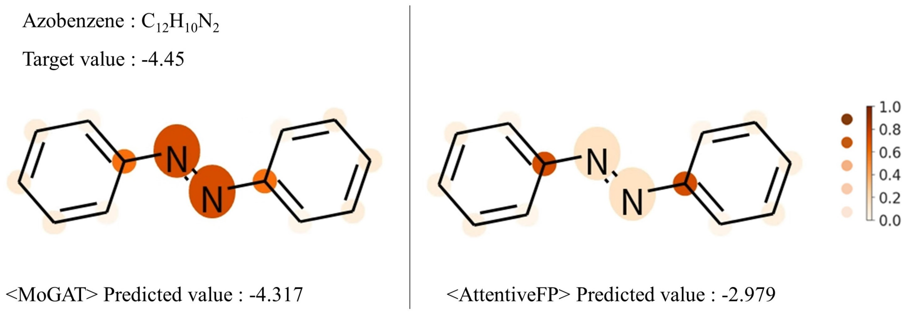

1-Bromonapthalene : $C_{10}H_7Br$

Target value : -4.35

<MoGAT> Predicted value : -4.132

<AttentiveFP> Predicted value : -4.383

**Figure S3.** The predicted water solubility (unit: logS) and importance of each atom for 1-bromonapthalene molecule. The attention scores are presented by the color coding in the right panel. In this case, the proposed method had a slightly larger prediction error than AttentiveFP; however, it shows better attention results than AttentiveFP.


Azobenzene : $C_{12}H_{10}N_2$

Target value : -4.45

<MoGAT> Predicted value : -4.317

<AttentiveFP> Predicted value : -2.979

**Figure S4.** The predicted water solubility (unit: logS) and importance of each atom for azobenzene molecule. The attention scores are presented by the color coding in the right panel.

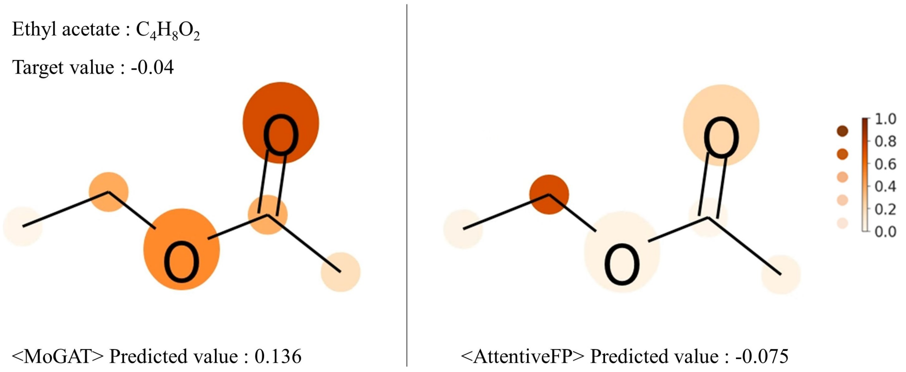

Ethyl acetate : C4H8O2
Target value : -0.04
O
O
O
O
1.0
0.8
0.6
0.4
0.2
0.0
<MoGAT> Predicted value : 0.136
<AttentiveFP> Predicted value : -0.075

**Figure S5.** The predicted water solubility (unit: logS) and importance of each atom for ethyl acetate molecule. The attention scores are presented by the color coding in the right panel. In this case, the proposed method had a slightly larger prediction error than AttentiveFP; however, it shows better attention results than AttentiveFP.

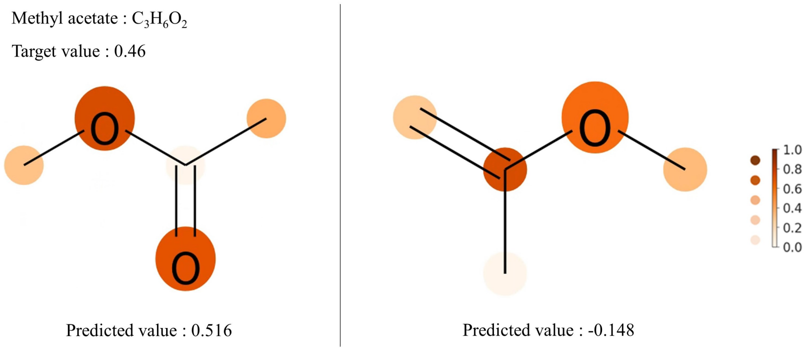


**Figure S6.** Change in water solubility and attention scores when one of oxygen atoms in methyl acetate is substituted with -$CH_2$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

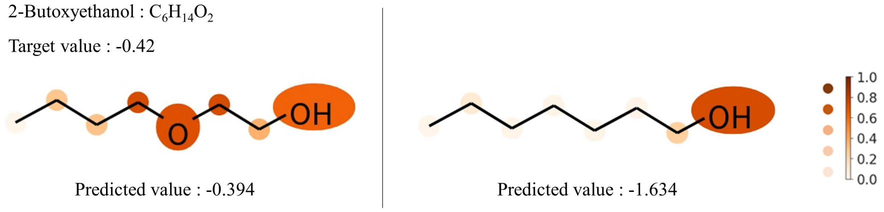


**Figure S7.** Change in water solubility and attention scores when one of oxygen atoms in 2-butoxyethanol is substituted with -$CH_2$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

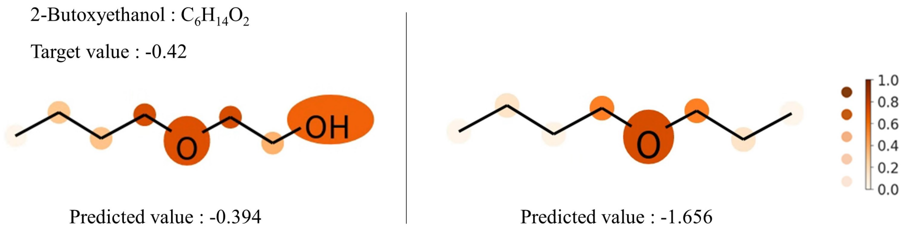


**Figure S8.** Change in water solubility and attention scores when an -OH in 2-butoxyethanol is substituted with -$CH_3$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

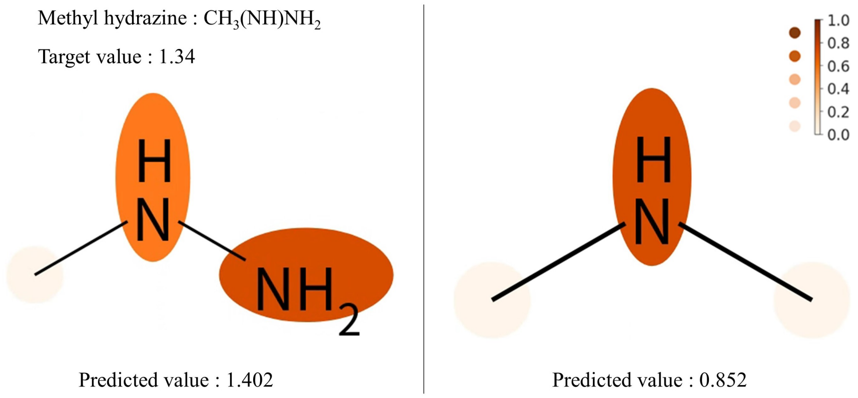


**Figure S9.** Change in water solubility and attention scores when a -NH2 in methyl hydrazine is substituted with $-CH_3$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

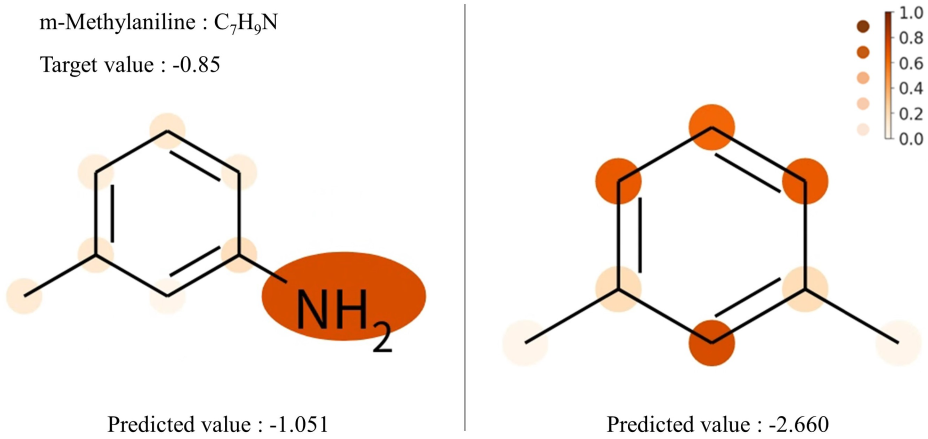


**Figure S10.** Change in water solubility and attention scores when a -NH2 in m-methylaniline is substituted with $-CH_3$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

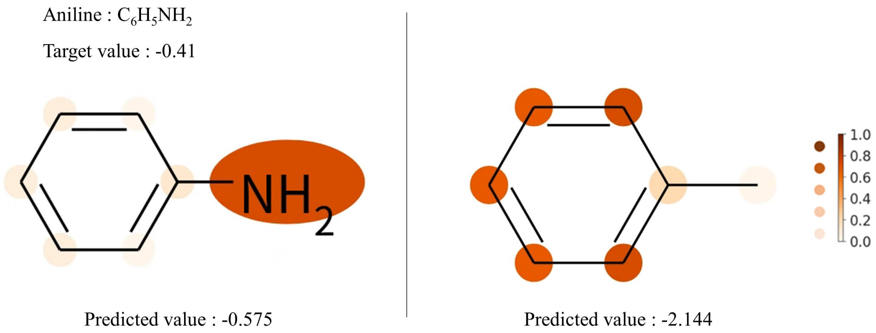


**Figure S11.** Change in water solubility and attention scores when a -NH2 in aniline is substituted with $-CH_3$. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

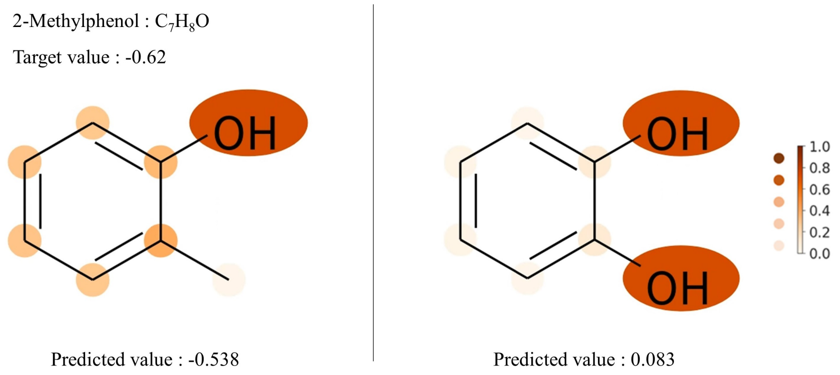


**Figure S12.** Change in water solubility and attention scores when a -$CH_3$ in 2-methylphenol is substituted with an -OH. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

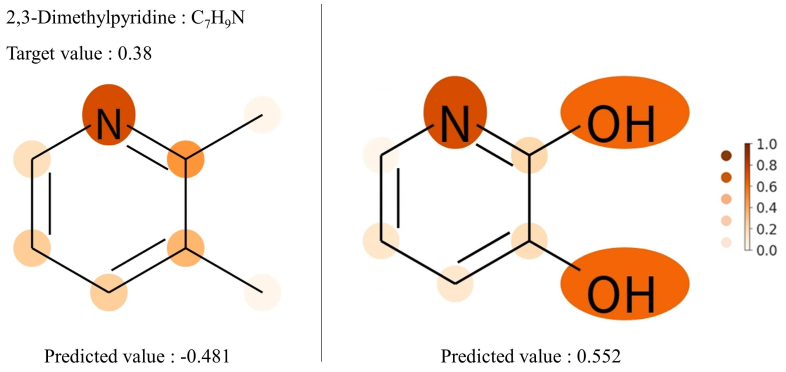


**Figure S13.** Change in water solubility and attention scores when -$CH_3$ in 2,3-dimethylpyridine are substituted with two -OHs. The unit of water solubility is logS and attention scores are presented by the color coding in the right panel.

## References

**1.** Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chem. science* **9**, 513–530 (2018).

**2.** Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *J. chemical information modeling* **52**, 2864–2875 (2012).