Electronic Supplementary Information

Design Green Chemicals by Predicting Vaporization Properties Using Explainable Graph Attention Networks

Yeonjoon Kim[a,b,†], Jaeyoung Cho[c,d,†], Hojin Jung[a,e], Lydia E. Meyer[c], Gina M. Fioroni[c], Christopher D. Stubbs[a], Keunhong Jeong[a], Robert L. McCormick[c], Peter C. St. John[c,*], Seonah Kim[a,c,*]

[a] *Department of Chemistry, Colorado State University, CO 80523, United States*

[b] *Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

[c] *National Renewable Energy Laboratory, 15013 Denver W Pkwy, Golden, CO 80401, United States*

[d] *Department of Aerospace and Mechanical Engineering, The University of Texas at El Paso, El Paso, TX 79968, United States*

[e] *Department of Chemical and Biomolecular Engineering, Yonsei University, Republic of Korea*

[†] *Equal contribution.*

[*] Corresponding Authors.

S1. Detailed information about split data sets

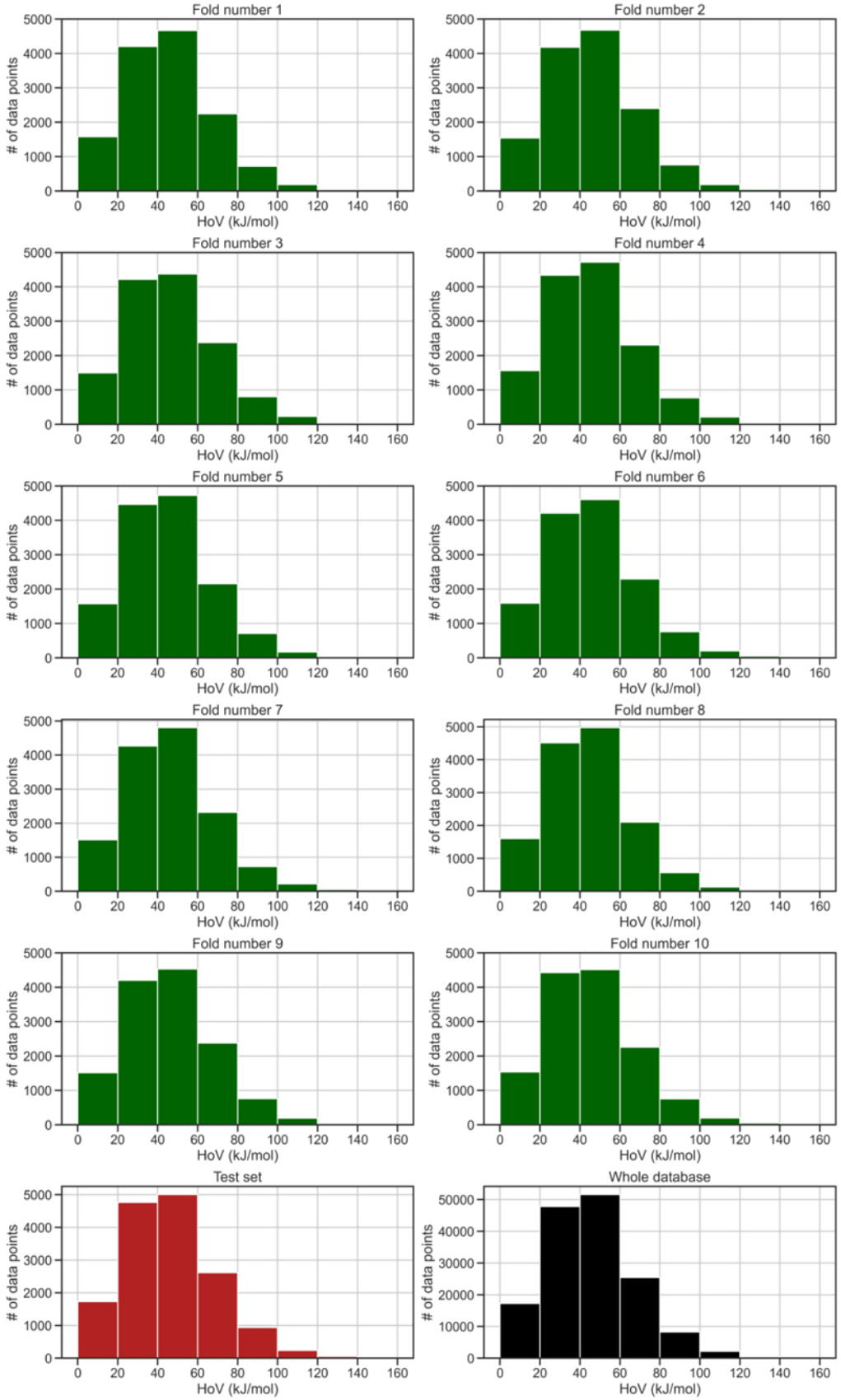


Fig. S1. Heat of vaporization distribution of each split data set.

Table S1. Detailed information about dataset splitting.

Dataset		$N_{molecule}$	N_{data}
Validation[a]	Fold 1	666 per each fold	13,634
	Fold 2		13,796
	Fold 3		13,561
	Fold 4		13,961
	Fold 5		13,828
	Fold 6		13,728
	Fold 7		13,916
	Fold 8		13,936
	Fold 9		13,615
	Fold 10		13,749
Test		740	15,371
Total		7,400	153,105

[a] For each fold, the remaining 5994 molecules from the other 9 folds (666 * 9) are training set molecules.

S2. Distribution of uncertainties

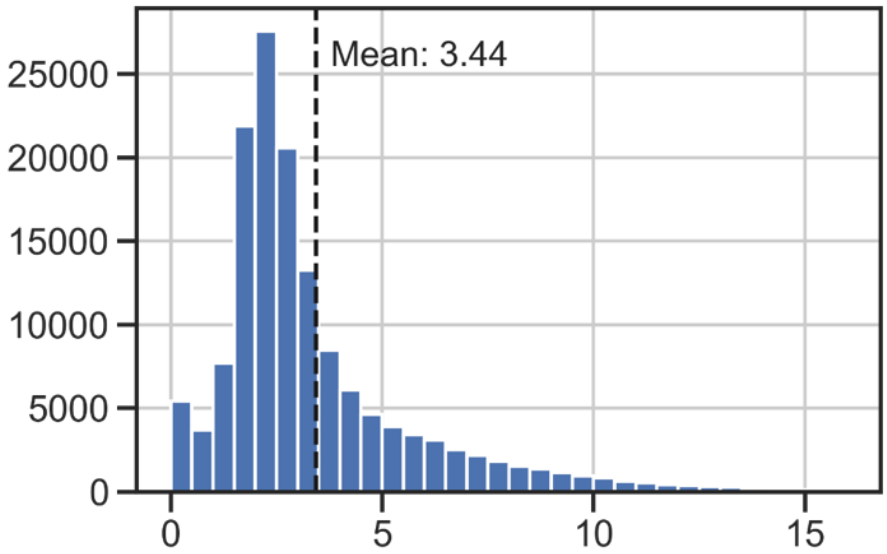


Fig. S2. Distribution of uncertainties of 153,105 heats of vaporization in the NIST WTT database. The vertical dotted line indicates the mean uncertainty (3.44 kJ/mol).

S3. Effect of the number of layers, attention heads, and loss function on model accuracy

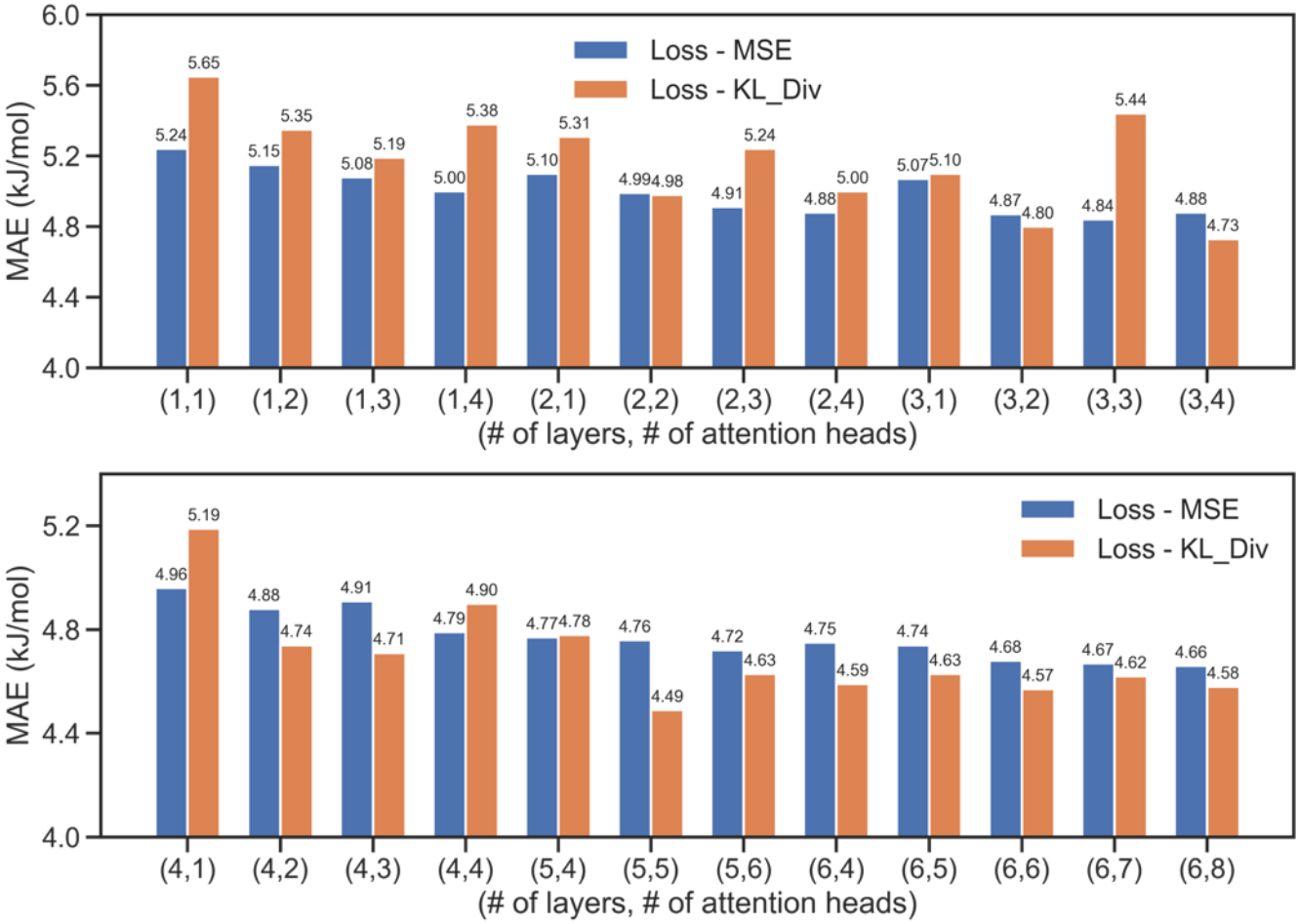


Fig. S3. Mean absolute errors from the 10-fold cross-validation of the models with different number of layers, attention heads, and loss functions.

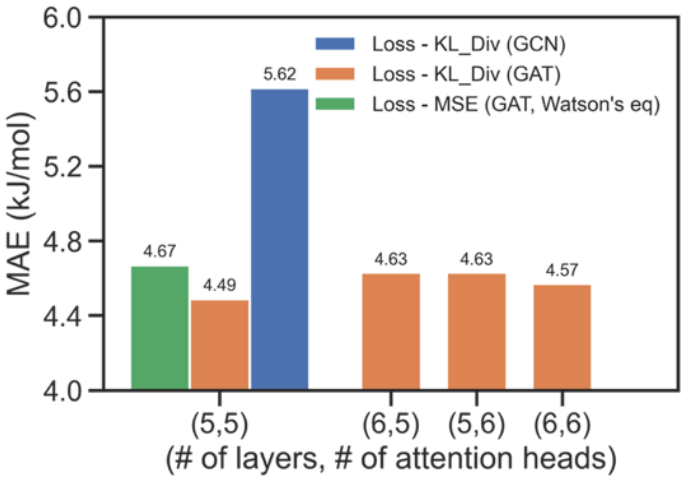


Fig. S4. Mean absolute error from the 10-fold cross-validation of the models with different graph neural network structures, and comparison with the equation-based model.

The prediction based on Watson's equation was carried out by obtaining three quantities (A, B, C) shown in Equation (S4), from the last readout layer from the graph attention networks depicted in Fig. 3a:

$$H_v = \frac{H_{vb}}{(T_c - Tb)^{0.38}}(T_c - T)^{C} \quad \text{(S4)}$$

A, B, and C values for each molecule were pre-determined by non-linear regression, and these values were trained using graph attention networks. Since the values A, B, and C have different orders of magnitude (e.g., 400<B<1600, 0<C<1), each one was scaled using the standard scaler:

Z = (X-□) / σ□□ □ □S□□□

□ and σ of A, B, and C were obtained from the training set molecules, implying that the prediction results can be biased by training sets. Training the reliable model without standardizing was not possible. Even though the Watson-equation-based model was trained under the 'privileged' conditions mentioned above (non-linear regression, standardization), it showed lower accuracy than the 'direct' prediction of HoV from the molecular structure. Moreover, 29 out of 7,400 molecules showed 'unphysical' values (i.e., A<0 or B<0 or C<0) in the non-linear regression, which makes it impossible to train or predict their HoVs if Watson equation is assumed.

S4. Optimization of other hyperparameters

Table S2. List of other hyperparameters tested.

Hyperparameter	Comments
Residual connection	'True' is better
Explicit hydrogen	'False' is better
Number of nodes in hidden layers	Using 32 is sufficient, no significant improvement when 64 was used
Dropout rate	0%, 5%, 10%, 15% were tested, applying no dropout is the best
(Learning rate, batch size)	(0.0005, 256), (0.01, 32) showed the best results for HoV and other properties, respectively
Epochs	200/500 is sufficient to reach the lowest training and validation set error for HoV/other properties

S5. Detailed analysis on the errors from HoV prediction

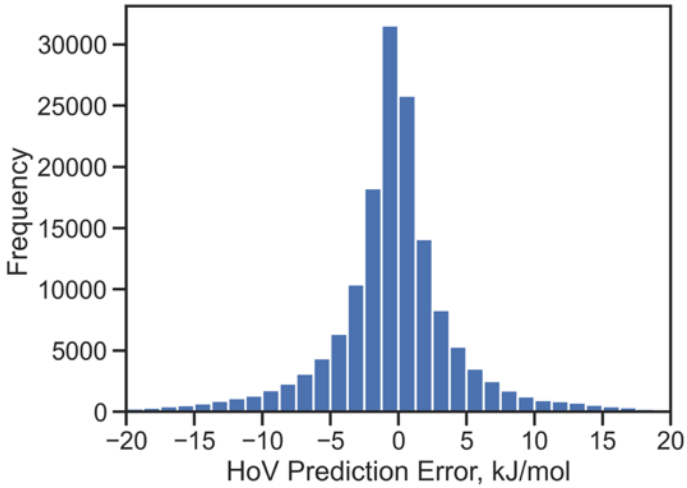


Fig. S5. A histogram plotting the distribution of HoV prediction errors.

Table S3. Mean absolute errors of HoV prediction by functional groups.

Oxygenates					Hydrocarbons				
Functional group	# of molecules	# of data points	# of data points/ molecule	Mean absolute error (kJ/mol)	Functional group	# of molecules	# of data points	# of data points/ molecule	Mean absolute error (kJ/mol)
Esters	1,044	21,932	21.01	3.12	Alkynes	153	3,350	21.90	2.66
Ethers	1,928	40,846	21.19	3.31	Alkenes	1,113	23,657	21.26	2.49
Carbonyls	2,088	43,541	20.85	3.89	Fused Rings	731	12,838	17.56	4.86
Alcohols	1,623	34,991	21.56	4.26	Cyclics	1,607	29,949	18.64	4.58
Peroxides	57	1,217	21.35	5.10	Alkane	454	9,708	21.38	2.24
Phenolics	331	6,477	19.57	4.61					
Furanics	35	706	20.17	2.24					
Water/CO	2	9	4.50	4.44					

[a] If more than two functional groups co-exist in a molecule, it is counted as a duplicate in each of the functional groups that the molecule contains.

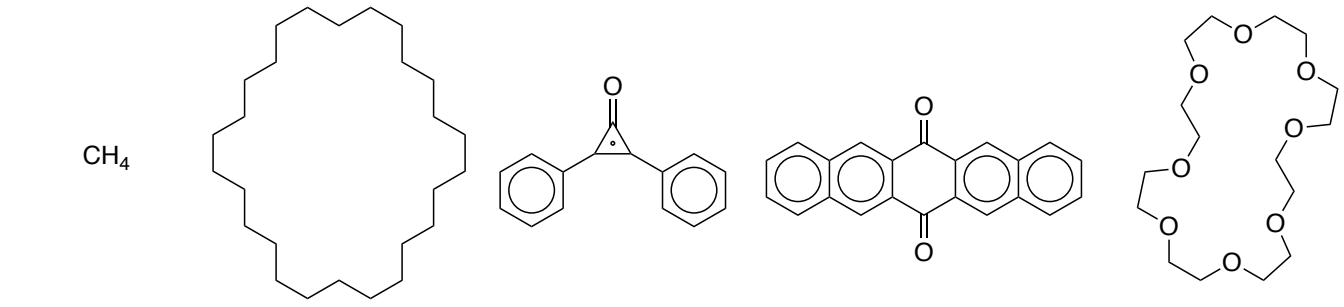


Fig. S6. Top 5 outliers of HoV prediction. Methane is an outlier because it is the only molecule containing carbon with four hydrogens, and there are only three data points. Including methane in the training set with more data points would significantly decrease the MAE (cf. water is in the training set – 5 data points, MAE of 5.92 kJ/mol). Other molecules have uncommon, peculiar structures. For example, 26-membered ring cycloalkane, 24-membered ring crown ether, a ketone containing the conjugation among one cyclopropane ring and two phenyl rings, and pentacene-like quinone structure.

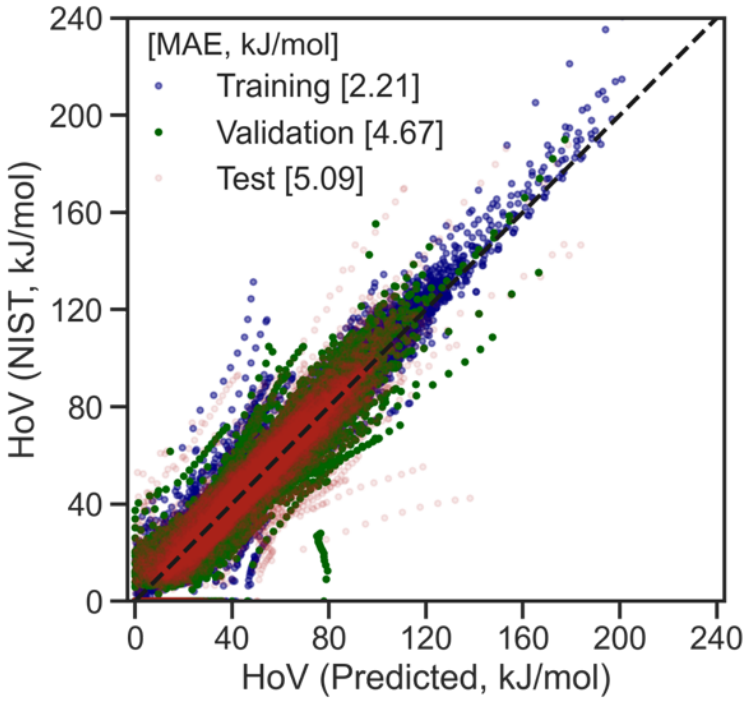


Fig. S7. Parity plot of the graph attention network model in which mean-squared-error loss function was used without uncertainty quantification.

Figure S8(a) counts the number of molecules containing each functional group in the HoV database, allowing the duplicated count if a molecule contains more than two functional groups. As shown, the HoV database from the present study for oxygenates is rich in carbonyls, ethers, alcohols, and esters, while those of phenolics, peroxides, and furanics are particularly underrepresented. Similarly, for the hydrocarbon, the number of molecules counts is highest for cyclic, which is followed by alkene, fused rings, and alkyne. This inhomogeneity of functional groups in the database influences the accuracy of the HoV prediction models, as described in Table S3. For reference, the functional group distribution in the other database – FP, T_B, T_C, T_M, and C_P – are depicted in Figure S8(b) – (f). Similarly to the HoV database, the molecules included in other properties' databases show a certain level of underrepresentation of some functional groups. For example, the FP database has a scarce amount of phenolics, peroxides, and furanics for oxygenates, while the number of hydrocarbon molecules is much smaller than those of oxygenates.

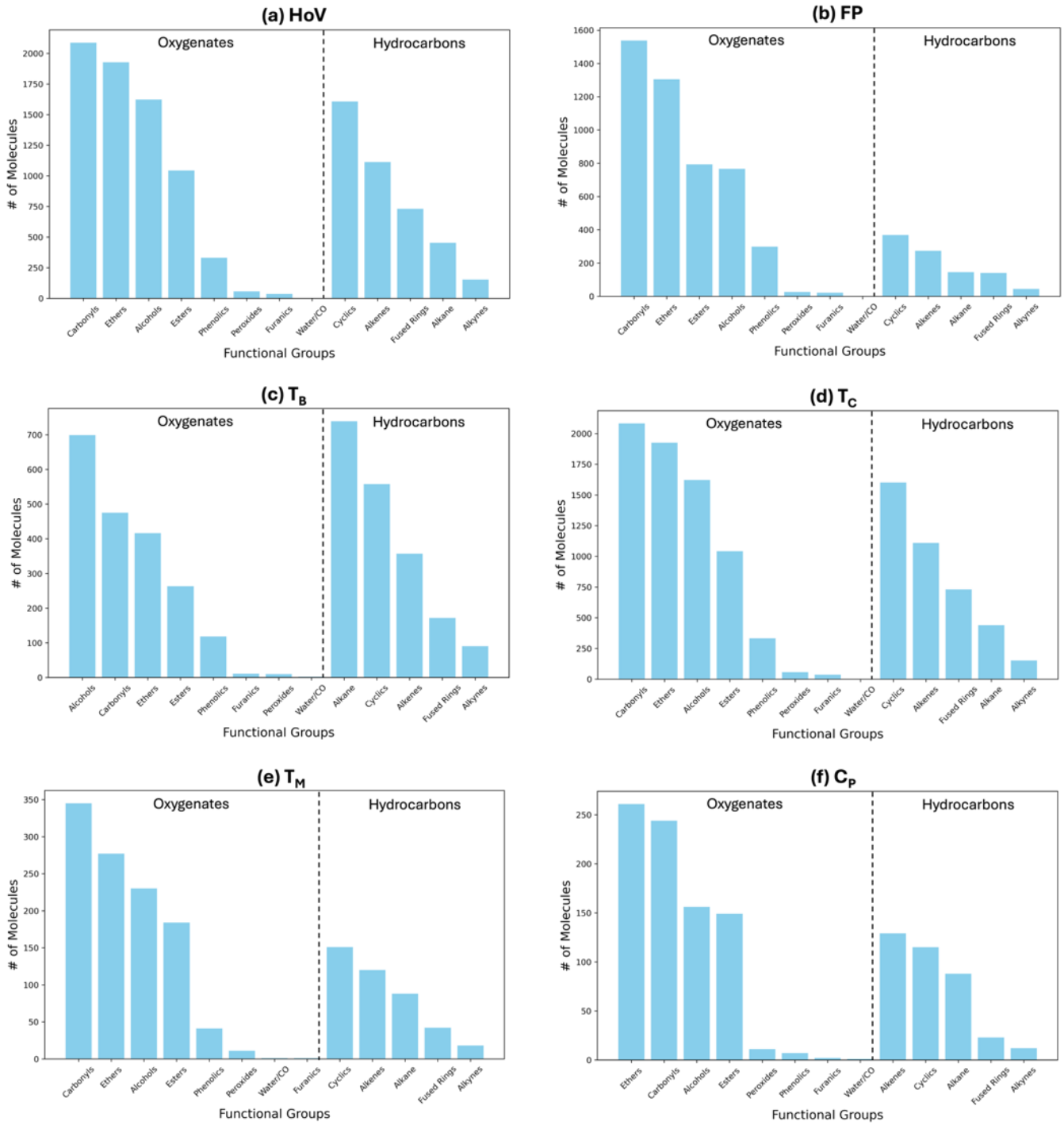


Fig. S8. The number of molecules containing each functional group in the HoV database

S6. Supplementary data for the flash point prediction

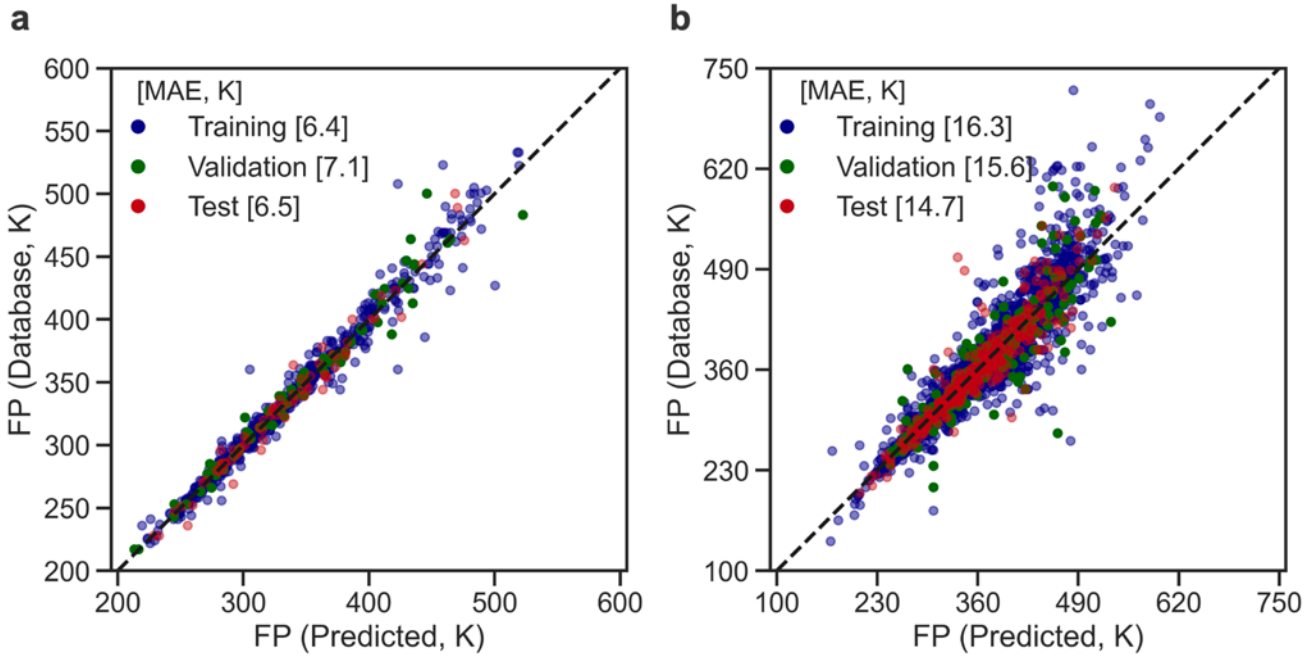


Fig. S9. Parity plot of the flash points predicted from the transfer learning model, for the (a) DIPPR database, and (b) the whole database obtained by collecting all the data from literature (3,282 data points – 2,626, 328, 328 for training, validation, and test set, respectively).

S7. Evaluation of atom-wise attention weights

The atom-wise attention of atom j ($\tilde{\alpha}_j$) in the 5th convolution layer was calculated by:

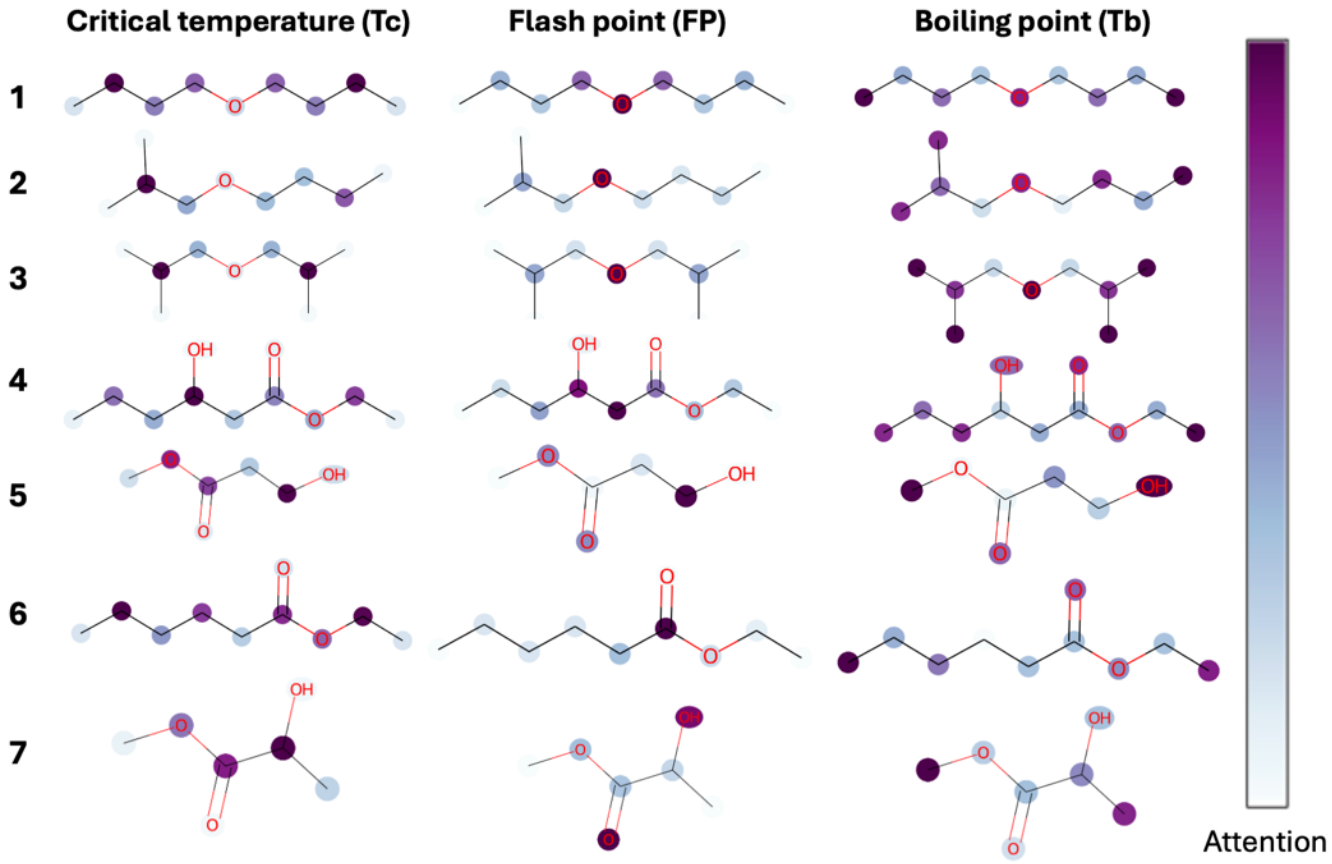
$$\tilde{\alpha}_j = \langle\alpha\rangle_j / \max(\langle\alpha\rangle_1, \ldots, \langle\alpha\rangle_N) \quad , \qquad \text{(S6)}$$

where

$$\langle\alpha\rangle_j = \sum_i^N \alpha_{ij}/N \ , \qquad \text{(S7)}$$

N is the number of atoms in a molecule, and a_{ij} is the element of attention matrix in Equation S1.

S8. Chemical interpretation of the models for predicting critical temperatures, flash points, and boiling points



	SMILES	Predicted_Tc	Exp_Tc	Predicted_FP	Exp_FP	Predicted_Tb	Exp_Tb
1	CCCCOCCCC	585.606	584.061	291.55927	298	412.7133	414.15
2	CC(C)COCCCC	570.77936	575	286.30545	295.95	402.6464	405.75
3	CC(C)COCC(C)C	559.0328	560.6	280.56543	287.65	392.33063	395.76
4	CCCC(O)CC(=O)OCC	665.1198	662	368.44415	367.15	489.76797	N/A
5	COC(=O)CCO	633.693	638	352.65305	345.75	457.44876	453.7
6	CCCCCC(=O)OCC	618.8482	615.2	326.5129	326	441.47186	441.2
7	COC(=O)C(O)C	598.5651	602	323.05725	324.15	414.51392	417.95

Fig. S10. Analysis of experimental/predicted critical temperatures (Tc), flash points (FP), boiling points (Tb), and atom-wise attention weights for three ethers and four esters.

Fig. S10 illustrates the attention weights obtained from the predictive models for critical temperatures (Tc), flash points (FP), and boiling points (Tb) with experimental and predicted property values, for three ethers and four esters. The trends of attention scores for Tc are analogous to those for HoVs (Fig. 6), which is consistent with the high relevance of Tc with HoV; Tc is the temperature where HoV becomes zero. A high attention score of an atom in FP prediction can be understood as having a high impact on the molecules' flammability and vaporization characteristics. For example, a high attention score for the oxygen atoms in the ether molecules is greatly consistent with the conventional knowledge, as the ether functional group plays an essential role in accelerating the low-temperature combustion by lowering the energy barrier of key reaction steps: the isomerization from the peroxy radical (RO_2) to the hydroperoxyl-alkyl radical (QOOH).[a] Meanwhile, ethyl hexanoate showed

the highest attention score at the carbon in the carbonyl site, which can be attributed to its pivotal role in disposing of carbon monoxide after the initial dissociation of esters.[b] Attention scores for the hydroxy-substituted esters – ethyl 3-hydroxyhexanoate, methyl 3-hydroxypropanoate, and methyl 2-hydroxypropanoate – were relatively harder to interpret from the existing knowledge, partially due to the limited literature on the combustion and vaporization characteristics of these distinct functional group. Still, the attention scores from the present study can be leveraged in future studies to understand how the molecular structures of hydroxyl-substituted esters affect their FP.

a. *Sustain. Energy Fuels,* 2022, 6, 3975-3988.
b. *J. Phys. Chem. A* 2023, 127, 9804-9819.

S9. Polymers that show the weak correlation between HoV and glass transition temperature

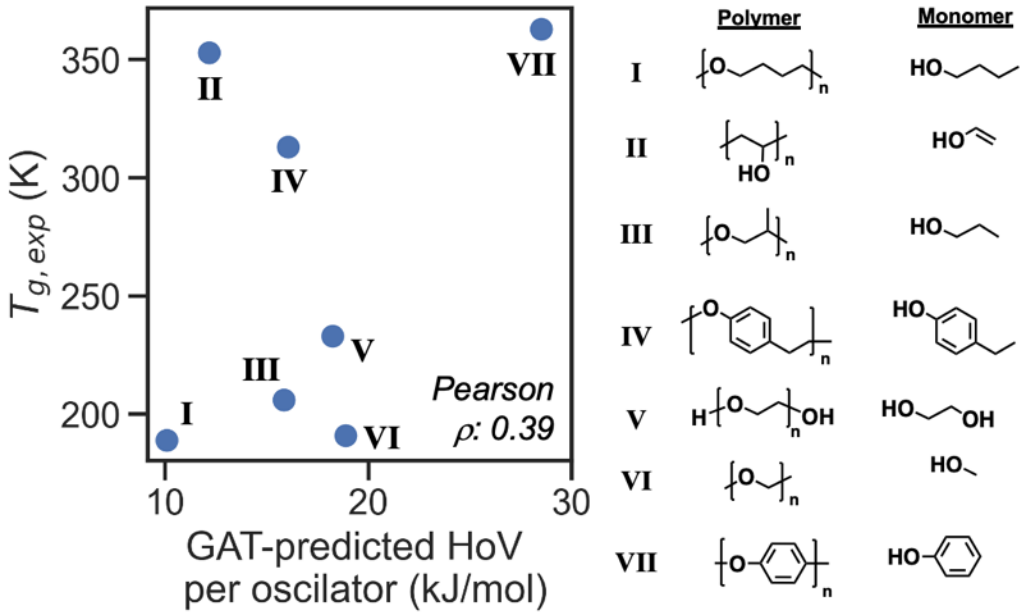


Fig. S11. Seven polymers that were omitted from the analysis shown in Fig. 9 (Predicting polymer's glass transition temperature from monomer HoVs) due to the weak correlation between HoV and glass transition temperature. Monomer structures of these polymers contain hydroxyl groups that cause errors in correlating monomer HoVs with glass transition temperatures.