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

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

[a]*Department of Chemistry, Colorado State University, Fort Collins, CO 80523, United States*
[b]*National Renewable Energy Laboratory, 15013 Denver W Pkwy, Golden, CO 80401, United States*
[c]*Department of Chemical and Biomolecular Engineering, Yonsei University, Republic of Korea*
[d]*Department of Mechanical Engineering, Colorado School of Mines, Golden, CO 80401, United States*

* *Corresponding author.* seonah.kim@colostate.edu
[†] *Equal contribution.*

ABSTRACT

Computational predictions of vaporization properties aid the *de novo* design of green chemicals, including clean alternative fuels and working fluids for efficient thermal energy recovery. Here, we developed chemically explainable graph attention networks to predict five physical properties pertinent to performance in utilizing renewable energy: heat of vaporization (HoV), critical temperature, flash point, boiling point, and liquid heat capacity. The predictive model for HoV was trained using ~150,000 data points, with considering their uncertainties and temperature dependence. Next, this model was expanded to the other properties through transfer learning to overcome the limitations due to fewer data points (700-7,500). Chemical interpretability of the model was then investigated, demonstrating that the model explains molecular structural effects on vaporization properties. Finally, the developed predictive models were applied to the design of chemicals that have desirable properties as efficient and green working fluids and fuels, enabling fast and accurate screening before experiments.

INTRODUCTION

Decarbonizing the power sector is one of the urgent missions for most countries to realize net-zero carbon emissions goal in the foreseeable future[1]. This will require advanced power generation technologies from renewable thermal resources (solar heat, geothermal, biomass, waste heat, etc.), necessitating the use of an efficient thermodynamic cycle that works in the low-to-mid temperature range. The organic Rankine cycle (ORC) has been recognized as a promising technology owing to its functionality over a wide temperature[2,3]. The ORC's performance heavily relies on the vaporization properties of organic working fluid[4]. For example, a working fluid with a high heat of vaporization (HoV) is known to give a higher unit work output at the given temperature of the heat source[5]. In this regard, extensive research has been conducted on the structure-property relationships for the working fluid's vaporization properties[6-9].

The vaporization properties of working fluids are also closely related to the performance of refrigerating cycles (or heat pumps)[10] that consume ~23 % of residential sector electricity in the United States[11]. Since the Montreal Protocol banned the use of chlorofluorocarbon, there have been constant demands for green working fluids with low global warming and ozone depletion potential[12]. Developing such chemicals must be preceded by a thorough understanding of structure-property relationships for vaporization properties.

The structure-property relationships of vaporization properties have been extensively studied for the purpose of designing clean (low-emission) alternative fuels[13-15]. Specifically, HoV has been considered one of the key factors determining the combustion characteristics of liquid fuels. Fuel vaporization in the engine cylinder leads to significant drop in temperature and pressure, affecting the thermal efficiency and emission characteristics of propulsion systems[16-18]. In this regard, a predictive model for particulate matter emissions from spark-ignition engines utilizes fuel HoV to account for the influence of its vaporization properties on the emission characteristics[19]. Similarly, the importance of HoV in the thermal efficiency of propulsion systems is evident as shown in the relationships of HoV vs. cetane number (CN)[20] and HoV vs. octane number (ON)[21].

A *de novo* design of green chemicals demands a predictive model for vaporization properties of arbitrary molecules. For HoV, various approaches have been applied to develop the predictive models, including the equation-based[22,23], group contribution (GC) models[24-26], and their combination with regression methods or neural networks[27-29]. Besides GC-based methods, quantitative structure-property relationship (QSPR) models have been built by using various structural descriptors[30-34]. Similar approaches have also been adopted for other vaporization properties[26,30,35-68], including critical temperature (T_C), flash point (FP), and boiling point (T_B).

Despite the remarkable advance in prediction accuracy over decades, these models still have several limitations. First, some of the equation-based models assume knowledge of prior information of other physical properties (e.g., T_B predictive equation as a function of HoV and vapor pressure). This assumption is sometimes problematic when one wants to assess a novel molecular structure whose physical properties have not been measured. Second, most models have not considered the temperature dependency of vaporization properties (e.g., HoV), which constrains the general applicability of the model to the broader temperature range. The majority of existing predictive models for HoV are valid for one temperature (room temperature or boiling point)[27-29,31,32]. Third, the models do not properly account for the uncertainties in experimental measurements. Training the model with uncertainty quantification can further improve model accuracy and provide confidence bound for the predicted value[69]. Lastly, there have been fewer discussions regarding chemical interpretation of predictive models than of their accuracy. A chemically explainable model can give the predicted value, and also the rational principles for designing green working fluids and low-emission fuels.

Here, we introduce a novel strategy to develop a reliable and chemically explainable machine learning (ML) predictive model for vaporization properties (**Fig. 1**). First, the databases of vaporization properties were collected and curated to use as inputs for training and evaluation of the model. The graph attention network (GAT) model was then built and trained against the databases. The model predicts the vaporization properties from a molecular graph where atoms and bonds are described as nodes and edges. It is an advanced graph neural network structure that can consider the effects of interactions among atoms on target molecular properties. Attention weights of each atom in GAT are related to structural importance and investigating them is beneficial in terms of their interpretability. Hence, it has been utilized in the prediction and analysis of numerous chemical properties[70-79].

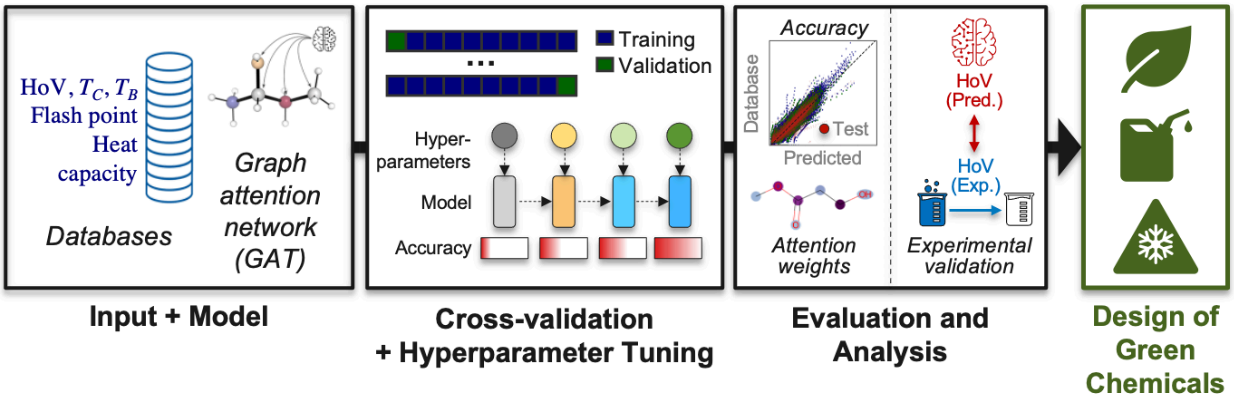


Fig. 1. Flow diagram of the overall procedure for developing predictive models for vaporization properties.

Besides GAT, tree-based ML algorithms have also been successful in various applications in chemistry, e.g., drug discovery[80]. However, in this work, we did not consider molecular descriptor-based models including tree-based ones because, first, our GAT showed better accuracy compared to the recent descriptor-based models (*vide infra*). Second, GAT does not usually need exhaustive molecular feature generation and selection. Reasonable accuracy was accomplished while using only minimal number of features (atom features and connectivity). Without incorporating additional molecular features, the model can infer overall molecular structural effects on HoV through local graph convolution which can consider more than first-nearest neighbors around each atom. Therefore, it could be generalizable to broader scope of molecules compared to descriptor-based models, and its accuracy can be comparable or better than conventional group contribution methods which usually consider only first-nearest atoms. Third, GAT is not computationally expensive when a graphical processing unit (GPU) is used. Details are available in the next sections, regarding the architecture and accuracy of the GAT model.

To reach the maximal accuracy, the optimal hyperparameters of the GAT were found by a grid search and ten-fold cross-validation. The mean absolute error (MAE) of validation sets from ten folds was evaluated for each hyperparameter, and the hyperparameter that showed the lowest MAE was selected. Among the ten models from the optimal hyperparameter set, the best model with the

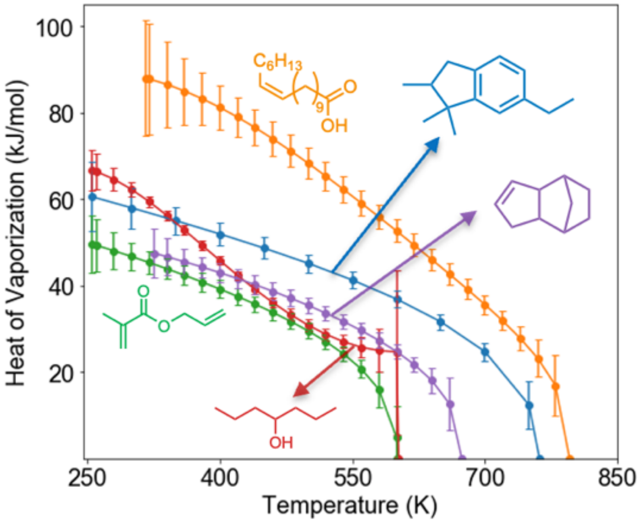


Fig. 2. Heat of vaporization of five example molecules in the NIST-WTT database.

lowest validation set MAE was selected. The final accuracy of the model with optimal hyperparameters was assessed for the held-out test set of HoV, with analyses of functional group effects and outliers. This training and accuracy evaluation process was then repeated for other properties: flash point (FP), critical temperature (T_C), boiling point (T_B), and heat capacity of liquid (C_P). The predictive model for HoV was also validated by comparing our experimentally measured HoVs with predicted values.

Subsequently, the chemical structural effects on HoV were investigated by analyzing the GAT model. Attention weights of each atom in a molecule were then compared to find key substructures or functional groups determining HoV. Such investigations demonstrate that our predictive model is accurate and chemically explainable. Finally, our predictive models for vaporization properties were applied to the practical design of green chemicals (i.e., working fluid and renewable fuel candidates). The following sections describe the detailed procedure and results obtained from each step outlined in **Fig. 1**.

Table 1. Summary of molecular properties and databases considered in this work.

Property	**N_{data}**	**References**	**Comments**
Heat of vaporization (HoV)	153,105	NIST Web Thermo Tables (NIST-WTT)[81]	• 7,400 molecules at different temperatures • Experimental + calculated values
Critical temperature (T_C)	7,362		• Temperature at which HoV is zero
Flash point (FP)	708	Design Institute for Physical Properties (DIPPR) database +Literature[27,29,31,32,46-48,50,52-54,56,82]	• 3,282 data points (DIPPR + literature). • Only 708 data points were used for training the model due to the inconsistency among different data sources.
Boiling point (T_B)	3,034		N/A
Heat capacity of liquid at 298 K (C_P)	777	DIPPR database[82]	• Control properties irrelevant to vaporization.
Melting point (T_M)	920		

RESULTS AND DISCUSSION

Databases of vaporization properties used for the model development

Table 1 summarizes the data sources and the number of data points for the six properties studied in this work. The present study only considers the molecules consisting of C, H, and O atoms, which are most common in fuels and working fluids readily synthesizable from natural sources. Halogens were omitted from the consideration owing to their potential impacts on ozone depletion. For the HoV prediction model, we used 153,105 data points of 7,400 molecules in the NIST Web Thermo Tables (NIST-WTT). **Fig. 2** illustrates the HoV values of five molecules in the NIST-WTT[81] as examples, depicting the sensitive nature of HoV to molecular structures. NIST-WTT contains the HoV values of each molecule at varying temperatures below T_C where HoV becomes zero. Of note, the database also provides error bars from experimental measurements or extrapolations from experimental values, which was utilized for uncertainty quantification of predicted HoVs. A tenth of the molecules (740 molecules) were reserved for the held-out test set for splitting the data. The rest 6,660 molecules were divided into ten folds to carry out the ten-fold cross-validation and hyperparameter tuning. Detailed information about each split data set is available in **Section S1** of **Supplementary Information**.

Meanwhile, T_C values of 7,362 molecules were collected from the same data source. FP of molecules were gathered from the Design Institute for Physical Properties (DIPPR) database[82] and other literature[57]. We removed the ambiguous FPs which are significantly different among multiple literature sources, leading to a total of 3,282 data points[46-48,50,52-54,56,82], 708 of which are from the DIPPR database. The FPs from the DIPPR database were only used for training the model, since combining all data from different sources deteriorates the predictive accuracy, presumably due to different reliability of standard and non-standard experimental methods (*vide infra* for details). The same procedure was repeated for T_B, resulting in 3,034 data points in total[27,29,31,32,82]. All T_B values correspond to those measured in the atmospheric pressure condition. In addition, 777 C_P

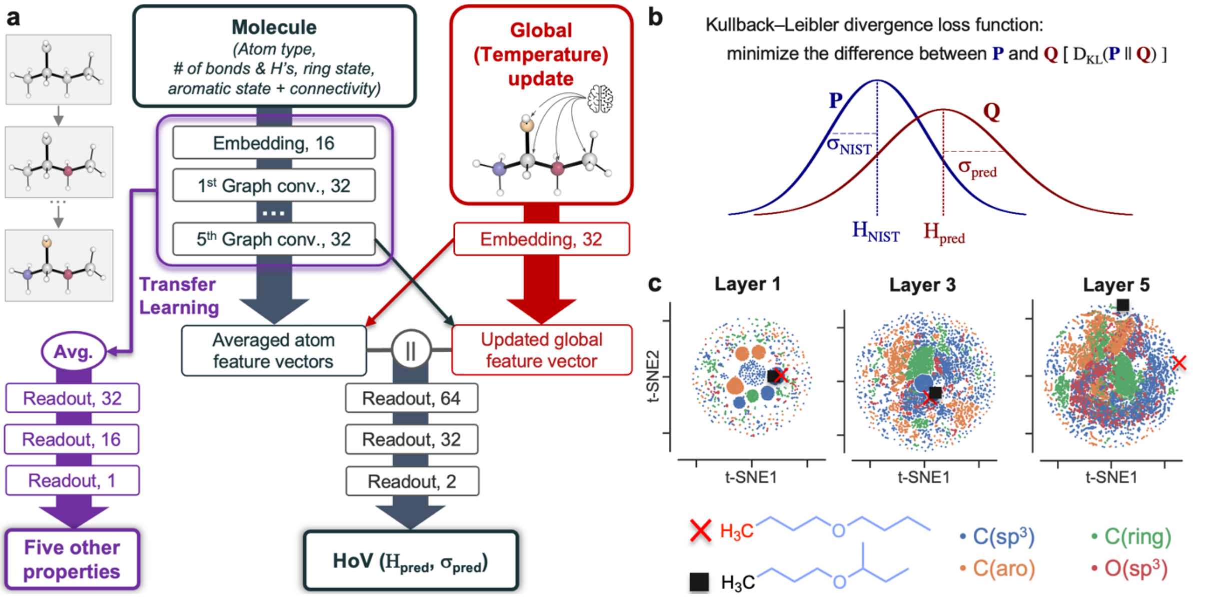


Fig. 3. (a) Architecture of the GAT model. (b) The Kullback-Leibler divergence loss function to predict HoV with considering uncertainty. (c) 2D representations of atom feature vectors obtained after passing the first (Layer 1), third (Layer 3), fifth (Layer 5) graph convolution layers. As a specific example, the feature vectors are plotted for two carbon atoms of dibutyl ether (in red cross) and butyl *sec*-butyl ether (in black square), to demonstrate that the model can consider the structural effect between an atom and its fifth-nearest neighbors.

values in liquid phase and 920 melting points (T_M) were acquired from the DIPPR database[82]. C_P and T_M were considered as a control group to compare the accuracy of predicting vaporization properties with those which are not related to vaporization. Of note, liquid C_P was also utilized with vaporization properties such as T_B, T_C, and HoV when designing new working fluids (*vide infra*).

Development of graph attention networks for predicting HoV

Fig. 3a shows a schematic diagram of our GAT model for predicting the HoV and other properties outlined in **Table 1**. The model first generates the 16-dimensional atom feature vectors from a simplified molecular-input line-entry system (SMILES) representation of a molecule. For each atom, five features are encoded as one-hot feature vectors. A connectivity matrix is also created from SMILES. These atom features and connectivity matrix comprise an input layer, and it should be emphasized that no three-dimensional coordinates of atoms in a molecule are needed for the prediction. Of note, SMILES strings can distinguish stereoisomers and diastereomers, and atom feature vectors can encode information about stereocenters. However, the current HoV model does not consider stereocenters, since only 13% of the molecules in NIST WTT contain the stereochemistry information (1,106 and 7,400 molecules with and without stereochemistry, respectively). In addition, the mean HoV difference between two stereoisomers (e.g., cis vs. trans, (E)- vs. (Z)-, and (R)- vs. (S)-) is 1.54 kJ/mol, being lower than the mean uncertainty of HoVs in NIST-WTT (3.44 kJ/mol, **Section S2** in **Supplementary Information**). Thorough consideration of stereochemistry effects on HoV is beyond the scope of current work and will be future work.

The input atom features then pass through the graph convolutional layers, and they are updated with consideration of neighboring atoms. Detailed formulations for graph convolution and attention matrices can be found in **Methods** and the literature[70]. Meanwhile, to consider temperature dependence on HoV, an input temperature value is embedded into a global feature vector. Next, the atom feature vectors from the last convolution layer are updated by the global feature vector, and those atom vectors again update the global feature vector (crossed arrows in **Fig. 3a**). More technical details about the global feature update scheme can be found in **Methods** and ref. [83]. The averaged atom feature vector and global vector are then concatenated and undergo three readout layers with ReLU activation functions to provide the predicted HoV (H_{pred}) and its uncertainty (σ_{pred}). In other words, the predicted HoV of a molecule is given as not a specific value, but a normal distribution **Q** whose mean and standard deviation is H_{pred} and σ_{pred}, respectively (**Fig. 3b**). This distribution is compared with another normal distribution $\mathbf{P}\sim N(H_{NIST}, \sigma^2_{NIST})$ acquired from the NIST database, and the model is trained so that the overlap between **P** and **Q** is maximized.

Methods for quantifying σ_{pred} include Bayesian neural networks (BNNs) where trainable weights and biases of readout layers are given as probability distributions instead of specific values. BNNs are appropriate for considering the epistemic uncertainty stemming from fitting the model to limited data. However, we assumed that the database is sufficiently extensive (153,105 data points, **Table 1**), and focused on aleatoric uncertainties arising from the variability from experimental measurements or extrapolation of experimental data. Such uncertainties may depend on uniquely complex molecular structures and can be irreducible regardless of database size[84]. In this regard, the final readout layer directly quantifies σ_{pred} as a function of molecular structure and outputs the distribution **Q**, instead of determining σ_{pred} from BNNs or ensembles of NNs. Elucidating the relationship between chemical structure and uncertainties informs how distant the molecule is located from the chemical space of well-known compounds and the fidelity of the predicted values when designing new molecules[85-88]. Recent studies have also adopted similar

Table 2. Comparison of accuracies of predicting HoVs with literature[a]

Reference	**Method**	**N_{data} (Training/ validation/ test)[a]**	**Mean absolute error (Training/ Validation/Test)**		**Comments**
			Literature	**This work (GAT)**	
Gharagheizi et al.[31]	Genetic algorithm-based multivariate regression	2291/ -/ 571	1.01/ -/ 0.99	0.66/ -/ 0.79	HoVs at boiling point (T_B)
Gharagheizi et al.[29]	Group contribution + artificial neural network	2312/ 287/ 275	0.86/ 1.21/ 1.05	0.84/ 1.20/ 1.16	HoVs at T_B
Jia et al.[32]	Features from quantum chemistry calculations + QSPR	219/ -/ 61	1.13/ -/ 1.12	0.88/ -/ 0.92	HoVs at T_B. Less extensive database but contains new oxygenates (alcohols, ethers, esters, ketones, etc.)

[a] C/H/O-containing molecules only. The model in the literature was trained using a larger database of organic molecules containing other elements.

approaches and obtained reliable results from the graph neural network-based prediction of molecular properties with uncertainty quantification[85,86].

In the first step of the model development, cross-validation and hyperparameter tuning were performed to find the best model architecture (**Fig. 1**). Using five layers with five attention heads minimizes the validation set MAE; fewer or more layers or attention heads do not improve the accuracy (**Section S3** in **Supplementary Information**). It should be noted that the mathematical definition of loss function is another key hyperparameter for developing a reliable model. The Kullback-Leibler (KL) divergence loss function, D_{KL}(P||Q), was adopted to minimize the difference between two normal distributions (**Fig. 3b**) of HoVs from the database and prediction. It has been successfully applied to recent ML models relevant to physics, chemistry, and biochemistry[89-92]. Detailed formula of the KL divergence is available in Equation (5) of the **Methods** section. Surprisingly, the KL divergence showed higher accuracy than the typical mean-squared-error loss function without uncertainty quantification, indicating that considering uncertainty is pivotal for a reliable prediction. In addition, the GAT model with the KL divergence is more accurate than the graph convolutional networks without attention and the GAT prediction based on Watson's equation (Details in **Section S3**, **Supplementary Information**. Optimization of other hyperparameters is explained in **Section S4** of **Supplementary Information**.

The weights of graph convolution layers from the HoV model were then used to expand the prediction to five other properties (**Fig. 3a**). A transfer learning approach was adopted to overcome the limitation due to fewer data points of these properties (700-7,500 data points, **Table 1**) compared to HoV (~10^5). Its feasibility was examined by comparing the accuracies of the models trained with and without transfer learning (For details, *vide infra*). These properties do not have temperature effect, so only the graph convolution layers were adopted from the HoV model. The averaged atom feature vectors obtained from the transfer learning pass through another series of readout layers to predict vaporization properties.

The five-layer GAT model (**Fig. 3a**) can distinguish the different local environments of atoms in a molecule, as shown in the t-stochastic neighbor embedding (t-SNE) analysis of atom feature vectors in hidden layers (**Fig. 3c**). The first layer's 2D t-SNE representations of atom features display a clear clustering according to the four basic atom types. Those in the third layer are more dispersed except for a few clusters near the center, and the fifth layer shows the most scattered atom features. This indicates that, as a molecular graph passes through more layers, the model updates atom feature vectors to differentiate more detailed local environments leading to different HoVs.

For further demonstration, we selected two representative compounds, butyl sec-butyl ether, and dibutyl ether which have slight structural differences in **Fig. 3c**. The former has one branched methyl group (methyl group on a tertiary carbon), whereas the latter does not. The terminal methyl carbons at the butyl group were chosen from each compound, and their atom feature vectors were compared. They show similar 2D t-SNEs until the third layer, and interestingly, they become distinct in the fifth layer. These two carbons share the same substructure until the fourth-nearest neighbors, but their fifth-nearest ones are different, and the model captures this structural dissimilarity, ultimately leading to different HoVs of these compounds.

The feasibility of the model shown in **Fig. 3a** was assessed by training the model using the databases of HoVs at T_B from the literature and comparing the prediction accuracies from previously reported models (**Table 2**). The previous studies used various techniques such as genetic algorithm, multivariate regression, group contribution, and artificial neural network. For fair comparison, we applied the splits of data sets into training, validation, and test sets identical with those reported in the literature. Although only C/H/O-containing molecules were chosen, the training:validation:test set ratio is maintained approximately to 8:1:1 (or training:test 4:1), which is reasonable for training our model and comparing the accuracy with other models. Our model

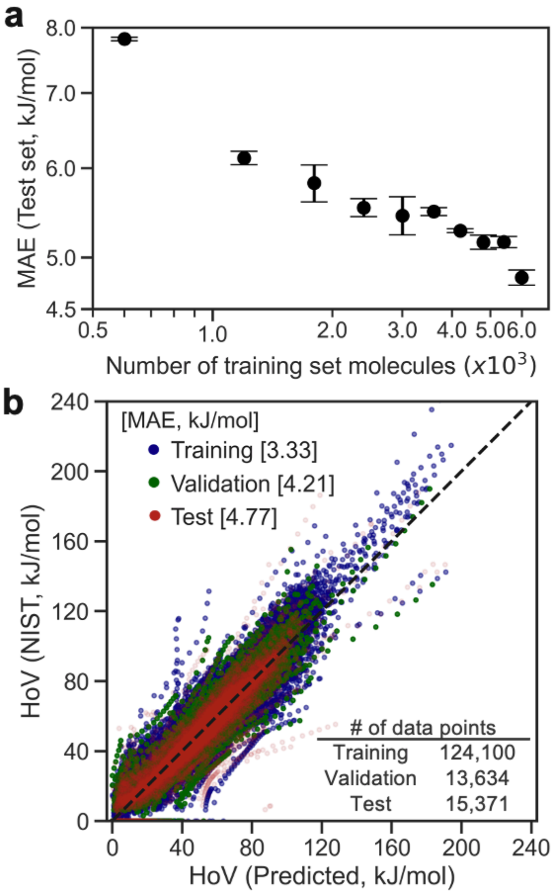


Fig. 4. (a) Learning curve for the model, plotting the test set MAEs against the number of molecules in the training set. Error bars indicate the standard deviation from triplicate runs. (b) Parity plot of predicted vs. database HoV values for training (blue), validation (green), and test (red) sets.

shows better accuracy in most cases, despite being trained using less extensive databases. A test set MAE 0.1 kJ/mol higher was shown in only one case, which could be attributed to experimental uncertainties or numerical noises from the model.

Accuracy of the HoV model trained using the largest database

Ultimately, our GAT model was trained using much more extensive database compared to any other models in literature. There are 124,100 HoVs at varying temperatures in the training, 13,634 in the validation, and 15,371 in the test sets. In the best-case model, we achieved reasonable accuracy for this large database, with the MAEs of 3.33, 4.21, and 4.77 kJ/mol for each split data set. Although the MAEs are relatively higher than those of HoVs at T_B (**Table 2**, 0.7-1.2 kJ/mol), it should be emphasized that the errors are comparable to the mean uncertainty of HoVs in the database (3.44 kJ/mol, **Section S2, Supplementary Information**). Given the MAEs similar to the database's mean uncertainty, it can be deduced that the GAT model architecture and the trained model is less susceptible to overfitting. Moreover, the model was trained using the largest database ever (153,105 data points) compared to any other previous studies, with taking temperature effects of HoV into account.

A learning curve was obtained (**Fig. 4a**) by training the model with increasing number of training set data points, where triplicate runs were performed for each training set to consider the variance of MAEs stemming from the randomness of training. A clear improvement of test set accuracy was shown as the number of training set molecules increased, suggesting that the model accuracy could be further improved by using a more extensive database.

More analysis on the model error was then carried out (Details in **Section S5**, **Supplementary Information**). The MAEs by 13 categorized functional groups were analyzed. All functional groups showed lower MAEs (2.24 – 4.57 kJ/mol) than the overall test set MAE (4.77 kJ/mol) except for fused ring compounds whose MAE is 5.03 kJ/mol. Fused rings have fewer number of data points per molecule at different temperatures (17.06 data points/molecule) than other functional groups (19-22 data points/molecule) while their structures are more complex, presumably leading to their higher MAE.

The molecular structure of top 5 outliers was further analyzed. Interestingly, methane showed the highest MAE (81.4 kJ/mol), which may be attributed to higher experimental uncertainty of HoV for molecules with low T_B (111 K for CH_4). The molecules with second to fifth highest MAE are complex cyclic compounds. The 2nd and 5th outliers have 26- and 24-membered ring, respectively, and their structures are highly twisted and deviated from typical conformations (chair and boat, etc.) of cyclic compounds. The remaining two compounds are cyclopropene with ketone and phenyl rings, and quinone with four linearly fused rings (pentacenequinone). Such structural distinctiveness is hard to be captured by GATs that use 2D structures as inputs, so they became outliers from predictions. However, these large-sized or fused ring structures are obviously uncommon and far from desirable fuel candidates or working fluids. To further examine the feasibility of uncertainty quantification, we compared the accuracy of this model with one that

Table 3. Correlations between absolute errors of prediction ($|H_{NIST} - H_{pred}|$) vs. uncertainties quantified from the model (σ_{pred}).

Dataset	$N_{molecule}$	N_{data}	Pearson ρ	Spearman ρ
Training	5,994	124,100	0.60	0.57
Validation	666	13,634	0.49	0.47
Test	740	15,371	0.54	0.50

used a mean-squared-error loss function without considering uncertainty. A lower training set MAE of 2.21 kJ/mol was observed, but validation and test set MAEs are 4.67 and 5.09 kJ/mol, respectively, indicating that overfitting occurs if uncertainty is not considered (**Section S5, Supplementary Information**).

Next, we investigated the Pearson and Spearman rank correlation coefficients (ρ) between the absolute errors from the prediction ($|H_{NIST} - H_{pred}|$) and uncertainties quantified from the model (σ_{pred}), as listed in **Table 3**. In principle, these two quantities should show a positive correlation; If the uncertainty is low, the prediction error should also be low. The KL divergence formula (Equation (5), **Methods** section) also well reflects this trend; the numerator and denominator contain $|H_{NIST} - H_{pred}|$ and σ_{pred}, respectively. A stronger positive correlation leads to the numerator and denominator being closer, and thus the minimization of divergence values. Meanwhile, the first term of Equation (5) prevents $|H_{NIST} - H_{pred}|$ and σ_{pred} from simultaneously diverging to infinity. The logarithm of the ratio between σ_{pred} and σ_{NIST} minimizes σ_{pred} so that it can be closer to the uncertainty tabulated in the database (σ_{NIST}).

A Pearson ρ close to 1 indicates that two variables have a relationship close to monotonic proportionality. A Spearman ρ equal to 1 corresponds to identical ranks of two variables. Our GAT model showed a decent positive Pearson correlation: 0.60, 0.49, and 0.54 for training, validation, and test set, respectively. The Spearman rank correlation values were located within 0.47-0.57. This is comparable to the ρ=0.469 obtained from the state-of-the-art message-passing neural network, which quantified the uncertainty for molecular properties of 133,885 compounds in the QM9 dataset[85]. All these results manifest that our model gives not only an accurate HoV prediction but also a reasonable quantification of uncertainties.

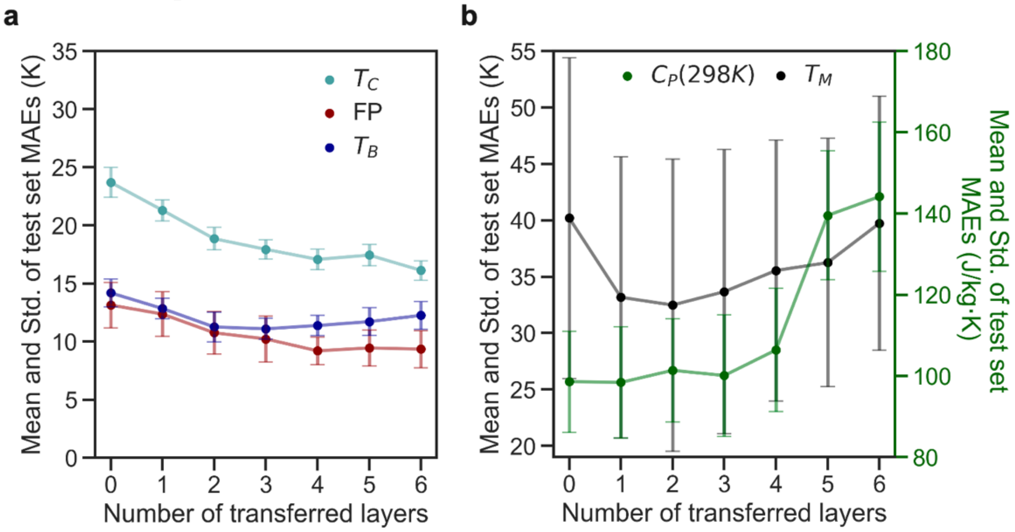


Fig. 5. The mean and standard deviation of test set MAEs of 20 GAT models from different random data splits, with varying the number of graph convolution layers transferred from the HoV model. Line and scatter plots with error bars for (a) three vaporization properties and (b) two properties irrelevant to vaporization.

Expansion of the predictive model to other vaporization properties

The predictive model for HoV was expanded to the prediction of other vaporization properties listed in **Table 1** by adopting the transfer learning approach (**Fig. 3a**). Such expansion is to overcome the limited number of data points for these properties, while utilizing the pre-trained HoV model that learned chemical structural effects on vaporization from the large database. Transfer learning can be carried out with varying the number of layers transferred from the HoV model. Here, we hypothesized that the relevance with HoV is different for each of the properties in **Table 1**, and transferring more layers is optimal when a property has higher relevance with HoV. For each property, the GAT models were trained with changing the number of transferred layers (0 to 6, seven cases) to find the optimal number of transferred layers and the model with the best accuracy. 20 different data set splits were tested for each of the seven cases to prevent the model from obtaining biased results regarding accuracies.

Fig. 5a illustrates the mean and standard deviation of test set MAEs from the 20 models for T_C, FP, and T_B with different numbers of transferred layers. The standard deviation of MAEs does not exceed 2 K for T_c, FP, and T_B, indicating that changing the data splits does not affect the overall trends of MAEs. These low deviations also demonstrate that the models from transfer learning are not susceptible to overfitting to specific data splits. These three vaporization properties are relevant to HoV, so it was effective to transfer all or part of the layers from the HoV model for maximizing the predictive accuracy. For T_C and FP, the means of test set MAEs converged with the difference below 1K, when four to six layers were transferred (16.1–17.1 K for T_C, 9.2–9.4 K for FP). Transferring two to five layers is optimal for T_B (Means of test set MAEs ranging from 11.1 to 11.7 K).

In contrast, C_P of liquid at 298 K and T_M are not related to HoV. These two properties were examined additionally for justifying that the optimal number of transferred layers is relevant to the relationship of a given property with HoV (**Fig. 5b**). Transferring 0–1 layers showed the best mean of test set MAEs (98.4–98.6 J/kg.K) for C_P. The optimal number of transferred layers is 1–2 for T_M, however, the means of MAEs (32–33 K) are much higher than those of other properties (9 – 17 K) shown in **Fig. 5a**. In addition, the standard deviations of MAEs are very high in all cases: 11 – 14K. These two contrastive examples further demonstrate our hypothesis that the number of transferred layers is related to the correlation between HoV and vaporization properties.

Table 4. Summary of the models for each vaporization property.

Property	**Number of transferred layers vs. correlation with HoV**			**Best-case models**		
	Number of transferred layers[a]	**Pearson coeff.**	**Correlation between**[b]	**N_{data} (Training/ Validation/Test)**	**MAE (Training/ Validation/Test)**	**Unit**
Critical temperature (T_C)	4 – 6	0.86	HoV at 298 K vs. T_C	(5,890/736/736)	(15.9/16.1/14.9)	K
Flash point (FP)	4 – 6	0.91	HoV at FP vs. FP	(566/71/71)	(6.4/7.1/6.5)	K
Boiling point (T_B)	2 – 5	0.68	HoV at T_B vs. T_B	(2,427/304/303)	(7.2/8.9/9.2)	K
Melting point (T_M)	1 – 2	0.18	HoV at T_M vs. T_M	(736/92/92)	(19.1/26.2/21.7)	K
Liquid heat capacity at 298 K(C_P)	0 – 1	-0.10	HoV at 298 K vs. C_P	(622/78/77)	(65.1/78.3/81.0)	J/kg.K

[a] Numbers of layers where the mean of test set MAEs is above within 1 K (1 J/kg.K for C_P) compared to the lowest one.
[b] HoVs are from the GAT predictive model and the target properties are from the database.

We also compared the Pearson correlation coefficient between HoVs and other vaporization properties (**Table 4**) to verify that a property is strongly correlated with HoV if the model becomes more accurate when more layers are transferred. The first target property is T_C; T_C is the temperature where HoV becomes zero. Watson's equation estimates that the HoVs at different temperatures T are proportional to $(T_C - T)^{22}$. In other words, there is a direct formulaic relationship between T_C and HoV, which can be associated with a high Pearson ρ (0.86) between HoV at room temperature and T_C. Transferring four to all six layers showed the best accuracy in predicting T_C, also being in line with these high Pearson ρ values. The Pearson ρ between FPs and HoVs at FP (0.91) is comparable to that in the case of T_C, resulting in the identical range of the optimal number of transferred layers (4–6 layers). Previous studies[45,51] quantified the relationship between FP and HoV. They derived an equation for estimating FP as a function of HoV, T_B, and other descriptors such as the number of carbons, surface area, etc., explaining the Pearson ρ value for FPs.

T_B is also known to have a relationship with HoV, according to the Clausius-Clapeyron equation and other studies regarding FP and T_B[45,51]. Therefore, transfer learning shows better accuracy than training the model without transferring layers from the HoV model, with slightly fewer number of transferred layers (2–5) than T_C and FP. It should be emphasized that the model for each vaporization property has been developed without the prior knowledge regarding the relationships among these properties, while the results are consistent with their underlying physical equations.

Meanwhile, the best-case model for each property should be chosen to use it for screening desirable working fluids and fuel candidates. **Table 4** summarizes the best-case models with their number of data points and MAEs for training, validation, and test sets. The best-case models showed the test set MAE of 14.9 K, 6.5 K, and 9.2 K for T_C, FP, and T_B, respectively. T_C could also be predicted by estimating the temperature where the predicted HoV becomes zero; however, the HoV prediction near T_C was less accurate than that at lower temperature ranges (**Fig. 4b**). As can be seen in **Fig. 2**, the uncertainties of NIST-WTT HoVs increase near T_C, leading to less reliable predictions of HoVs when they approach to zero. To obtain the best T_C prediction accuracy, transfer learning was carried out, instead of predictions from the HoV model, resulting in the best model shown in **Table 4**.

The FP prediction model was developed using only the DIPPR database. Of note, we also attempted to train the model using a larger integrated database, but the MAEs increased (**Section S6**, **Supplementary Information**). The less accuracy for the larger database is presumably due to the inconsistency arising from different data sources including FPs measured using non-standard methods[46-50,52-56,82], rather than the deficiency of the model. The best model from training against the DIPPR database showed the MAEs of 6.4-7.1 K for training, validation, and test sets. These errors are comparable to the typical experimental errors of FP measurements using standard methods (5.0-8.0 K)[57,81,82]. On the other hand, the model for T_M showed a higher test set MAE (21.7 K) than other properties, but it was not used for designing green chemicals. The lowest MAEs for C_P of liquids are 65–81 J/kg.K. This accuracy is acceptable to be utilized in the design of working fluids (*vide infra*).

While numerous models have been reported for 'one independent predictive model per one property', all these results manifest the general applicability of the temperature dependence of HoV to other properties relevant to vaporization. Such approaches would lead to robust predictive models that are consistent with the underlying physics of vaporization and are integrated in one model architecture. The model can be more powerful if it is chemically interpretable, which is discussed in the next section.

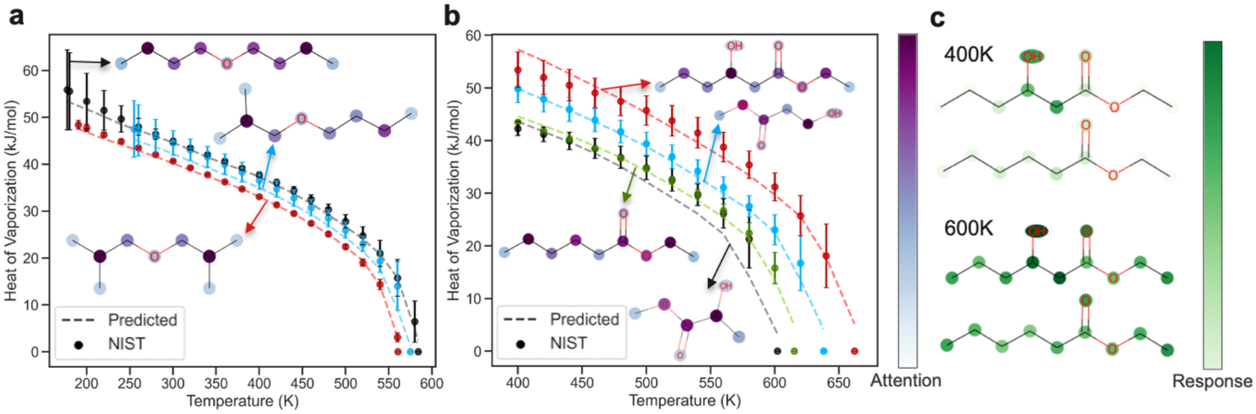


Fig. 6. Analysis of HoVs and atom attention weights for (a) three ethers: dibutyl ether (black), butyl isobutyl ether (blue), diisobutyl ether (red), and (b) four esters: ethyl 3-hydroxyhexanoate (red), ethyl hexanoate (green), methyl 3-hydroxypropanoate (blue), methyl 2-hydroxypropanoate (black). (c) Comparison of temperature response of atom feature vectors in ethyl 3-hydroxyhexanoate and ethyl hexanoate, at two temperatures.

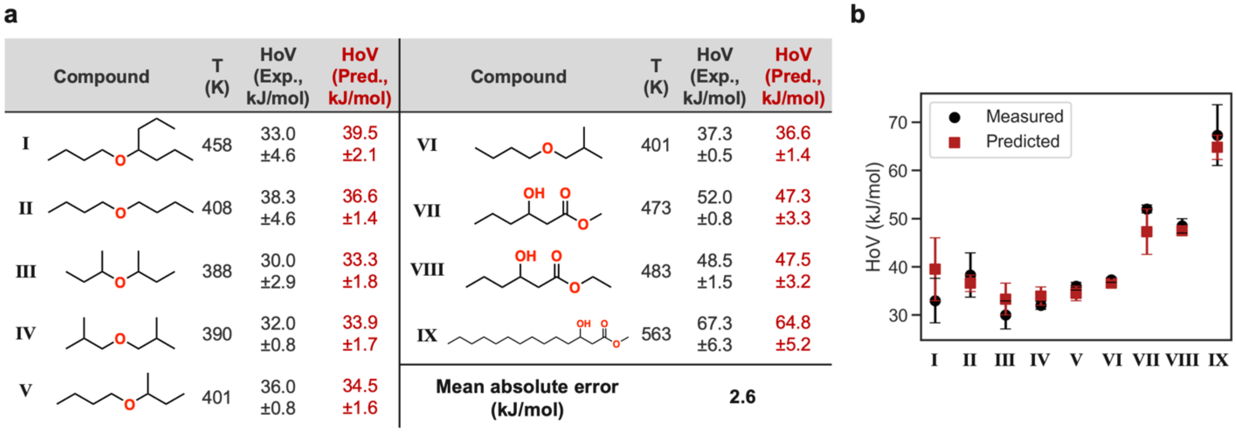
Chemical interpretation of the model

Interpretability of an accurate predictive model is a key aspect of the chemistry-informed design.[93,94] To demonstrate our model's chemical interpretation, we chose ethers and esters as representative molecules among various fuel candidates. They have drawn attention as promising biofuel candidates due to their favorable reactivity, emission characteristics, and viability of their synthesis from biomass[95,96]. First, the attention weights of atoms were analyzed to find the key substructures that lead to HoV differences. The literature[97] and **Section S7** in **Supplementary Information** explain the detailed procedure for evaluating atom-wise attention weights.

The attention weight analysis for three C_8 ethers is illustrated in **Fig. 6a**. The predicted HoVs showed a good agreement with those in the NIST-WTT. More methyl branches result in lower HoVs (dibutyl ether > butyl isobutyl ether > diisobutyl ether), presumably due to the decrease in molar surface area and, thus, intermolecular interactions[98]. The attention weights also explain this trend; the highest attention weights were observed in the tertiary carbons of two branched ethers since they have methyl branches and lower HoV than a linear one. The γ carbons in dibutyl ether showed the highest attention because they are adjacent to terminal methyl carbons and determine the continuation or termination of alkyl chains.

This analysis was repeated for esters (**Fig. 6b**). The hydroxy (OH) substitution at beta carbon of ethyl 3-hydroxyhexanoate (E3OHH) leads to higher HoVs than ethyl hexanoate (EH) because it can form intramolecular and intermolecular hydrogen bonds. HoVs of the hydroxyester with a shorter carbon chain (methyl 3-hydroxypropanoate: M3OHP) are still higher than EH, indicating the significance of OH groups in determining HoV. Our model also captured this structural feature; the beta carbons having an OH group showed the highest attention weights among atoms in E3OHH and M3OHP. On the other hand, the effect of OH position on HoVs was investigated. The HoVs of methyl 2-hydroxypropanoate (M2OHP) are lower than M3OHP. In both cases, the carbon atom with an OH group showed the highest attention, regardless of whether it is a terminal carbon or not.

The OH group also influences the temperature dependence of a molecule on HoV. For example, the HoV difference between E3OHH and EH at 600K is higher than that at 400K, indicating that E3OHH is more resistant to temperature change than EH. To verify this, we compared the response of atom feature vectors to the global updates, which is evaluated by the L2-norm of feature vector



Compound	T (K)	HoV (Exp., kJ/mol)	HoV (Pred., kJ/mol)	Compound	T (K)	HoV (Exp., kJ/mol)	HoV (Pred., kJ/mol)
I	458	33.0 ±4.6	39.5 ±2.1	VI	401	37.3 ±0.5	36.6 ±1.4
II	408	38.3 ±4.6	36.6 ±1.4	VII	473	52.0 ±0.8	47.3 ±3.3
III	388	30.0 ±2.9	33.3 ±1.8	VIII	483	48.5 ±1.5	47.5 ±3.2
IV	390	32.0 ±0.8	33.9 ±1.7	IX	563	67.3 ±6.3	64.8 ±5.2
V	401	36.0 ±0.8	34.5 ±1.6	Mean absolute error (kJ/mol)		2.6	

Fig. 7. (a) Results from our in-house measurements of HoVs for nine ether and hydroxy ester molecules, with HoV values predicted using our GAT model. (b) Overlapped confidence intervals of measured and predicted HoV values for these nine molecules.

difference before and after the update: **||v-v'||** (**Equation 2** in **Methods** and **Fig. 6c**). At 400K, all atoms in EH and E3OHH show a low response value to the temperature except the OH group, alpha, and beta carbons of E3OHH. The overall responses increase at 600K, but these three atoms in E3OHH respond most sensitively to the external temperature, leading to a slower HoV decrease of E3OHH than EH as temperature increases. This indicates that the OH substitution at beta position is a key factor for increasing the HoV of esters.

The above analysis on attention weights and temperature dependence demonstrates our model's capability of capturing chemical structural effects on HoV. The predicted HoVs are accurate and are consistent with the chemical knowledge pertinent to HoV, such as molecular surface area and hydrogen bonds. The structural insights gained from this chemical interpretation would inform the discovery and design of new working fluids and (bio)fuel candidates. It should be emphasized that the chemical interpretation method using attention weights can also be applied to the GAT models trained through transfer learning for other vaporization properties.

Experimental validation of the model

We carried out in-house measurements of HoVs at temperatures near T_B for further assessment of the model using the external data besides NIST-WTT. HoVs were measured for three beta-hydroxy esters and six ethers shown in **Fig. 7a**. They are promising biofuel candidates derivable from biomass and have high reactivity and low soot emission[95,96,99]. They also have diverse structural features such as linear/branched, symmetric/asymmetric alkyl chains, hydroxy, ether, and ester groups, which is good for model evaluation. Particularly, three of them (4-butoxyheptane, methyl 3-hydroxyhexanoate, and methyl 3-hydroxytetradecanoate: **I**, **VII**, and **IX**) do not exist in NIST-WTT. The rest six compounds are found in NIST-WTT, but the GAT model has never seen HoVs at the temperatures given in **Fig. 7a** during the model training. Therefore, the feasibility of our external validation is further justified by the unavailability of these nine molecules at the given temperatures.

We predicted HoVs of these molecules at the same temperature using our model and compared the values from the measurement and prediction. As a result, our GAT model showed reasonable accuracy with an MAE of 2.6 kJ/mol for these nine molecules. It should be emphasized that all measured and predicted values overlap if uncertainties are considered (**Fig. 7b**), which manifests the importance of considering confidence intervals in the ML prediction of HoV.

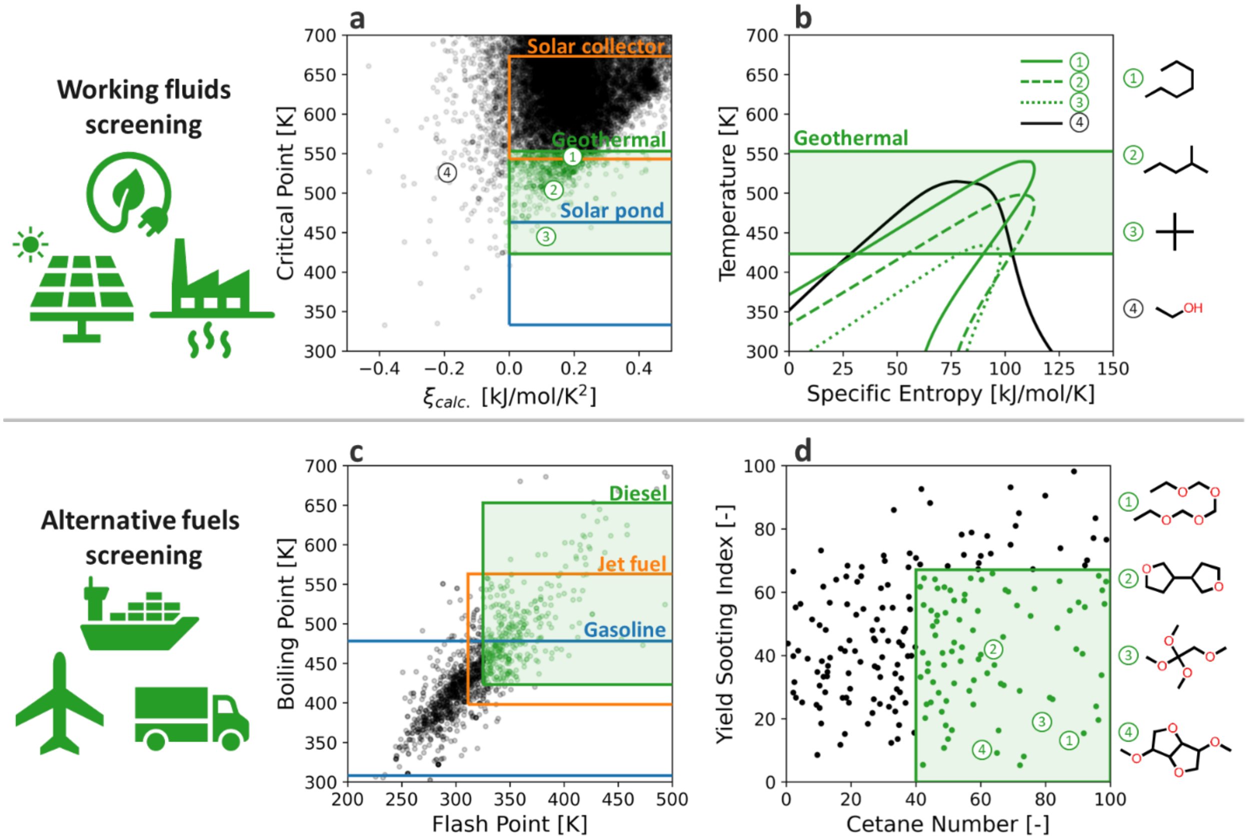


Fig. 8. Application of the GAT model for working fluid and alternative fuels screening. (a) Distribution of ~27,000 organic molecules on T_C - ξ axis, (b) T-s curve of four different working fluids with varying T_C and ξ, (c) distribution of ~1,300 saturated ethers on T_B – FP axis, and (d) sub-screening based on YSI and CN.

Application of the model to green chemical screening

The developed GAT models for vaporization properties prediction can have numerous potential applications for designing green chemicals. One example is to screen out the optimal ORC working fluids with desirable vaporization properties that maximize the utility of renewable thermal resources. Xu et al.[100] discussed the relevance of working fluids' T_C on the thermal efficiency of sub-critical pressure ORC. Their simulation study revealed that the thermal efficiency of ORC at a given temperature of heat source (T_H) is maximized with the working fluids having T_C in between of $T_H – 30$ K and $T_H + 100$ K, suggesting the T_C as an essential criterion for screening the optimal working fluids. Meanwhile, the "dryness" of working fluids was also widely accepted as an important property relevant to ORC's thermal efficiency and work output[101-103]. The working fluid is considered dry if upon isentropic expansion of the saturated vapor, the fluid stays in the vapor phase, which is essential to ensure the absence of liquid droplets at the turbine exit. The dryness of the working fluid can be determined with the temperature sensitivity of the specific entropy ($\xi = ds/dT$) of saturated vapors; that is, the working fluid is dry if $\xi>0$ or wet otherwise. Liu et al.[101] suggested an analytic equation for predicting ξ of organic compounds from their vaporization characteristics as below:

$$\xi_{calc.} = \frac{1}{T_H^2}\left(C_{P,l}T_H - \left(\left(\frac{nT_H^*}{1-T_H^*}\right) + 1\right)\mathrm{HoV_H}\right), \qquad (1)$$

where T_H^* is the reduced temperature of the heat source ($=T_H/T_C$), n is an empirical coefficient that ranges from 0.375 to 0.38[104], and $\mathrm{HoV_H}$ is the HoV at T_H. This study assumes the T_H as T_B for the brevity in molecular screening.

To screen out the working fluids based on their dryness and T_C, **Fig. 8a** depicts the distribution of ~27,000 organic molecules from the database (NIST WTT, DIPPR, PubChem, etc.[81,82,105]) on T_C – ξ axis, where all the relevant molecular properties – T_C, T_B, $C_{P,l}$, and $\mathrm{HoV_H}$ – were evaluated from the present GAT model. The T_C screening criteria for solar collector, geothermal, and solar pond were based on their typical temperature range (573 K, 453 K, and 353 K, respectively[106]), while ξ was restricted to positive. The majority (96 %) of tested molecules fall into the dry working fluid. Meanwhile, there are more compounds at higher T_C, providing more viable options for working fluid selection for high-temperature heat sources such as a solar collector. On the other hand, the low-temperature heat sources (geothermal and solar ponds) have limited choices for the dry working fluid.

The validity of working fluid screening based on the GAT model was confirmed on the *T*-*s* diagram of the selected working fluids for geothermal ORC (**Fig. 8b**), where the thermodynamic properties of liquid-vapor transition were collected from CoolProp[107]. The *n*-heptane met the screening criteria as a working fluid for geothermal ORC, and its *T*-*s* diagram in Fig. 7b depicts the ideal shape in geothermal temperature with clear dryness, proving the soundness of ML-based screening of ORC working fluid. Similarly, the *iso*-hexane and *neo*-pentane also satisfied the screening criteria for geothermal ORC but with lower T_C than *n*-heptane, which is consistent with their *T*-*s* diagram in **Fig. 8b**. This finding is in line with previous studies on *n*-heptane, *iso*-hexane, and *neo*-pentane as ORC working fluids[108,109], all of which showed a plausible performance in geothermal power generation. As a counterexample, we depicted the *T*-*s* diagram of ethanol, which shows the negative temperature sensitivity of specific entropy (thus, wet) as predicted from the ML-based working fluid screening. In summary, the GAT model from the present study can provide useful guidance on screening ORC working fluid for renewable thermal resources with varying temperatures.

As another example, the present GAT model can be utilized to discover alternative fuel candidates for decarbonizing the transportation sections. Our previous study[99] suggested ether fuels as a promising alternative to conventional fuels owing to their high reactivity and low soot emission characteristics while being synthesizable from biomass conversion. Despite the extensive studies from both experimental and theoretical approaches, the optimal structure of ether-containing molecules is still under investigation owing to the variety of their structural degrees of freedom. In this regard, the present study examined the utility of the GAT model in screening out ether fuels based on their vaporization and combustion characteristics.

ASTM standards[110-112] restrict various molecular properties of transportation fuels to ensure safety and operability in the propulsion systems. T_B range is one of the important criteria for categorizing the fuel molecules into diesel, jet fuels, and gasoline, and it affects the vaporization process of the injected fuels in the combustion chamber. Meanwhile, the safety of fuel and its inflammability is controlled by regulating the FP above certain criteria. **Fig. 8c** presents the distribution of ~1,300 saturated ethers on T_B – FP axis, where both properties are predicted from the GAT model from the present study. We set the boundary of T_B for diesel, jet fuel, and gasoline as 423 – 653 K, 398 – 563 K, and 308 – 473 K, respectively[113]. The lower limit of FP of diesel and jet fuels were set as 325 K and 311 K, while those of gasoline are not constrained, as described in ASTM standards. Consequently, 30.3 % of tested ethers fall into the diesel regime, while 45.3 % and 78.5 % are in the jet-fuels, and gasoline range, respectively. Of note, the currently oxygenated compounds such as ethers are not acceptable as alternatives to conventional jet fuels owing to their poor thermal stability and low specific energy[113]. Therefore, here we focused on diesel fuel candidates, although it can also be applied to the design of renewable fuels for any other engines including gasoline and aviation.

The 387 diesel-range ethers were then further analyzed on the cetane number (CN) and yield sooting index (YSI) axis, which represents the reactivity and sooting tendency of fuel candidates, as shown in **Fig. 8d**. The CN and YSI of ether compounds were estimated from the multivariate linear regression model suggested by Cho et al.[99]. The screening criteria for CN was set to be higher than 40 as dictated in ASTM standard for diesel fuels[111], while YSI was assumed below those of *n*-dodecane (YSI = 67.1), which is a typical surrogate fuel for conventional diesel. Consequently, the 60 of 387 diesel-range ethers satisfied the criteria for combustion characteristics. Figure 8d shows four of the selected ethers fuels, all of which contain multiple oxygen atoms to increase the reactivity and suppress the soot formation, as envisioned by Cho et al.[99]. In summary, the GAT model from the present study can provide an additional window for screening out the alternative fuels based on their vaporization characteristics, which significantly reduces the efforts in the combustion properties characterization.

CONCLUSIONS

The GAT model was developed for predicting vaporization properties. The extensive HoV database consisting of ~150,000 data points were utilized for the model development considering the temperature dependence of HoV and uncertainty quantification. The model showed a good prediction accuracy with reasonable uncertainty estimation. The predictive model for HoV expanded to other vaporization properties, whose databases are less extensive than HoV. It was beneficial to adopt transfer learning approaches for T_C, FP, and T_B, where the trained layer weights from the HoV model are used. The transfer learning models showed lower errors in estimation of these properties than the models from non-transfer training. The prediction and chemical interpretation were possible by analyzing attention weights and temperature response of atom feature vectors, leading to elucidation of molecular structural effects on HoV. Such workflow encompassing uncertainty quantification, transfer learning, and chemical interpretation was applied to the practical design of working fluids and (bio)fuel candidates. The computational approaches introduced in this contribution can be applied to other molecular properties related to the design of green chemicals, facilitating the clean and sustainable energy production.

METHODS

Our GAT model was programmed in Python 3.7[114] using the Deep Graph Library 0.7[115] with the TensorFlow 2.4[116] backend. In the GAT, the given 16-dimensional input features $H^{(0)}$ pass through graph convolution layers considering attention weights (a) that impose different convolution weights to each bond based on different surrounding atoms. The updated atom feature vector of atom i at the $(l+1)$-$^{\text{th}}$ layer [$H_i^{(l+1)}$] is:

$$H_i^{(l+1)} = \tau \left[\frac{1}{K} \left(\sum_k^{K} \sum_{j \in N(i)} \alpha_{ij,k}^{(l)} H_j^{(l)} W^{(l)} \right) \right], \qquad (2)$$

where τ is the rectified linear unit (ReLU) activation function to introduce non-linearity between molecular structure and predicted HoV, K is the number of attention heads. $N(i)$ is the set of first-nearest neighbors of atom i connected by bonds, $W^{(l)}$ is a graph convolution matrix. a and $W^{(l)}$ are trainable matrices.

The first update is carried out by:

$$\mathbf{v}' = \mathbf{v} + \tau[\phi_1(\mathbf{v}) + \phi_2(\mathbf{u})], \qquad (3)$$

where **v** and **v**' are the atom feature vectors before and after the update, respectively. **u** is the global (temperature) feature vector. ϕ_1 and ϕ_2 are two fully connected layers, respectively. The second update is performed by using the averaged atom feature vectors:

$$\boldsymbol{u}' = \boldsymbol{u} + \tau\left[\phi_3\left(\frac{1}{N_{atom}}\sum_i^{N_{atom}} v_i'\right) + \phi_4(\boldsymbol{u})\right], \qquad (4)$$

where **u** and **u'** are the global feature vectors before and after the update, respectively. ϕ_3 and ϕ_4 are two dense layers. v_i' is the updated feature vector of one atom obtained from Equation (2), and N_{atom} is the number of atoms in a molecule.
The KL divergence is defined as

$$D_{KL}(\mathrm{P}||\mathrm{Q}) = \frac{1}{N_{\mathrm{data}}}\left[\sum_i^{N_{\mathrm{data}}}\left(\log\frac{\sigma_{\mathrm{pred},i}}{\sigma_{\mathrm{NIST},i}} + \frac{\sigma^2_{\mathrm{NIST},i}+\left(\mathrm{H}_{\mathrm{NIST},i}-\mathrm{H}_{\mathrm{pred},i}\right)^2}{2\sigma^2_{\mathrm{pred},i}} - \frac{1}{2}\right)\right] \qquad (5)$$

where $\mathrm{H}_{\mathrm{NIST}}$, σ_{NIST}, $\mathrm{H}_{\mathrm{pred}}$, σ_{pred} are HoVs and uncertainty from database and prediction, respectively, and $\mathrm{P}\sim N(\mathrm{H}_{\mathrm{NIST}}, \sigma^2_{\mathrm{NIST}})$, $\mathrm{Q}\sim N(\mathrm{H}_{\mathrm{pred}}, \sigma^2_{\mathrm{pred}})$. It took about two hours to train the HoV model against 153,105 data points for 200 epochs using one V100 GPU.

Experimental details of HoV measurements.
Pure component symmetric ethers and beta hydroxy hexanoate esters investigated for HoV measurement were purchased in >98% purity from Sigma Aldrich. Asymmetric ethers were custom synthesized by Advanced Molecular Technologies of Melbourne, Australia. A Differential Scanning Calorimeter/Thermogravimetric Analyzer (DSC/TGA) (TA Instruments, Q600-series) was utilized to perform HoV measurements for comparison to predictive values and was based on a previous method developed for gasoline samples[117,118]. The instrument was calibrated per manufacturer's specifications, and a correction factor was calculated for the instrument (1.17) using n-butyl benzene because its HoV is well documented[119,120]. Utilizing a similar methodology to that developed by Luning Prak and coworkers[121], each pure component was placed in an aluminum pan (TA Instruments, Tzero Pan 901683.901) with a hermetically sealed pinhole lid (TA Instruments, Tzero Hermetic Lid w/ Pin Hole 901685.901). The DSC/TGA was held isothermally for one minute and then ramped at a rate of 30°C per minute until it reached a temperature 15-20°C below the boiling point of the pure component. The DSC/TGA was then held isothermally for 30 seconds, before again being ramped at a rate of 10°C per minute until it reached a temperature within 5°C of the boiling point. It then remained isothermal until the sample had completely evaporated as determined by the TGA. The heat flow was integrated from the start of the isothermal ramp until the end of the sample evaporation. The HoV was calculated as the

integrated heat flow divided by the mass loss as recorded by the TGA. Each sample was run in triplicate, and the average HoV was reported.

DATA AVAILABILITY

Subscriptions are required to access the property data in NIST-WTT and DIPPR databases. The molecules used for training the model are available through a GitHub repository (https://github.com/BioE-KimLab/HoVpred) with their property values from NIST-WTT and DIPPR redacted. The data points from literature were not redacted and are available through the GitHub repository.

CODE AVAILABILITY

The Python codes, list of molecules, and trained model files are available through a GitHub repository (https://github.com/BioE-KimLab/HoVpred).

ACKNOWLEDGMENTS

We acknowledge Chris Muzny and Demian Riccardi (NIST) for their help with access to NIST-WTT. This research was funded by the ExxonMobil Corporation. The views represented are those of the authors, and do not necessarily reflect those of ExxonMobil. The computer time was provided by the NSF Extreme Science and Engineering Discovery Environment (XSEDE), Grant no. TG-CHE210034 and by the National Renewable Energy Laboratory Computational Science Center. This work was authored in part by Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. H. Jung acknowledges the support by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0008747) supervised by the Korea Institute for Advancement of Technology (KIAT).

AUTHOR CONTRIBUTIONS

Y. K., J. C., R. L. M., P. C. S. J., and S. K. did conceptualization of the manuscript. Y. K. developed the Python code for ML models of vaporization properties. Y. K. and H. J. did the database curation of vaporization properties. L. E. M., G. M. F., and R. L. M. carried out the HoV experiments. K. J. performed 10-fold cross-validation and hyperparameter tuning of the model. J. C. applied the ML models to the design of working fluids and fuels and analyzed the results. All authors contributed to preparing and reviewing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information: Supplementary Sections S1-S7, Figs. S1-S7, Tables S1-S3.
Correspondence: Correspondence to Seonah Kim (seonah.kim@colostate.edu).

References

1 Bistline, J. E. & Blanford, G. J. The role of the power sector in net-zero energy systems. *Energy and Climate Change* **2**, 100045 (2021).

2 Loni, R. *et al.* A review of solar-driven organic Rankine cycles: Recent challenges and future outlook. *Renew. Sust. Energy Rev.* **150**, 111410 (2021).

3 Haghighi, A., Pakatchian, M. R., Assad, M. E. H., Duy, V. N. & Alhuyi Nazari, M. A review on geothermal Organic Rankine cycles: modeling and optimization. *J. Therm. Anal. Calorim.* **144**, 1799-1814 (2021).

4 Bao, J. & Zhao, L. A review of working fluid and expander selections for organic Rankine cycle. *Renew. Sust. Energy Rev.* **24**, 325-342 (2013).

5 Chen, H., Goswami, D. Y. & Stefanakos, E. K. A review of thermodynamic cycles and working fluids for the conversion of low-grade heat. *Renew. Sust. Energy Rev.* **14**, 3059-3067 (2010).

6 Papadopoulos, A. I., Stijepovic, M. & Linke, P. On the systematic design and selection of optimal working fluids for Organic Rankine Cycles. *Appl. Therm. Eng.* **30**, 760-769 (2010).

7 Su, W., Zhao, L. & Deng, S. Simultaneous working fluids design and cycle optimization for Organic Rankine cycle using group contribution model. *Appl. Energy* **202**, 618-627 (2017).

8 Peng, Y., Su, W., Zhou, N. & Zhao, L. How to evaluate the performance of sub-critical Organic Rankine Cycle from key properties of working fluids by group contribution methods? *Energy Convers. Manag.* **221**, 113204 (2020).

9 Luo, X. *et al.* Improved correlations for working fluid properties prediction and their application in performance evaluation of sub-critical Organic Rankine Cycle. *Energy* **174**, 122-137 (2019).

10 Piña-Martinez, A., Lasala, S., Privat, R., Falk, V. & Jaubert, J.-N. Design of Promising Working Fluids for Emergent Combined Cooling, Heating, and Power (CCHP) Systems. *ACS Sustainable Chem. Eng.* **9**, 11807-11824 (2021).

11 Annual Energy Outlook 2022., https://www.eia.gov/outlooks/aeo/ (March 3, 2022) (Energy Information Administration, 2022).

12 Choudhari, C. & Sapali, S. Performance investigation of natural refrigerant R290 as a substitute to R22 in refrigeration systems. *Energy Procedia* **109**, 346-352 (2017).

13 St. John, P. C., Kim, S. & McCormick, R. L. Development of a Data-Derived Sooting Index Including Oxygen-Containing Fuel Components. *Energy Fuels* **33**, 10290-10296 (2019).

14 Liu, H., Yoo, K. H., Boehman, A. L. & Zheng, Z. Experimental Study of Autoignition Characteristics of the Ethanol Effect on Biodiesel/n-Heptane Blend in a Motored Engine and a Constant-Volume Combustion Chamber. *Energy Fuels* **32**, 1884-1892, doi:10.1021/acs.energyfuels.7b03726 (2018).

15 Hulwan, D. B. & Joshi, S. V. Performance, emission and combustion characteristic of a multicylinder DI diesel engine running on diesel–ethanol–biodiesel blends of high ethanol content. *Appl. Energy* **88**, 5042-5055, doi:https://doi.org/10.1016/j.apenergy.2011.07.008 (2011).

16 Huang, Y., Li, Y., Luo, K. & Wang, J. Biodiesel/butanol blends as a pure biofuel excluding fossil fuels: Effects on diesel engine combustion, performance, and emission characteristics. *Proc. Inst. Mech. Eng. D* **234**, 2988-3000, doi:10.1177/0954407020916989 (2020).

17 Yang, P.-M. *et al.* Comparison of carbonyl compound emissions from a diesel engine generator fueled with blends of n-butanol, biodiesel and diesel. *Energy* **90**, 266-273, doi:https://doi.org/10.1016/j.energy.2015.06.070 (2015).

18 Ratcliff, M. A. *et al.* Impact of ethanol blending into gasoline on aromatic compound evaporation and particle emissions from a gasoline direct injection engine. *Appl. Energy* **250**, 1618-1631, doi:https://doi.org/10.1016/j.apenergy.2019.05.030 (2019).

19 St. John, P. C., Kim, S. & McCormick, R. L. Development of a Data-Derived Sooting Index Including Oxygen-Containing Fuel Components. *Energy & Fuels* **33**, 10290-10296, doi:10.1021/acs.energyfuels.9b02458 (2019).

20 Mishra, S., Anand, K. & Mehta, P. S. Predicting the Cetane Number of Biodiesel Fuels from Their Fatty Acid Methyl Ester Composition. *Energy Fuels* **30**, 10425-10434, doi:10.1021/acs.energyfuels.6b01343 (2016).

21 Wang, C., Zeraati-Rezaei, S., Xiang, L. & Xu, H. Ethanol blends in spark ignition engines: RON, octane-added value, cooling effect, compression ratio, and potential engine efficiency gain. *Appl. Energy* **191**, 603-619, doi:https://doi.org/10.1016/j.apenergy.2017.01.081 (2017).

22 Watson, K. M. Thermodynamics of the Liquid State-Generalized Prediction of Properties. *Ind. Eng. Chem.* **35**, 398-406 (1943).

23 Morgan, D. L. Use of transformed correlations to help screen and populate properties within databanks. *Fluid Ph. Equilibr.* **256**, 54-61, doi:https://doi.org/10.1016/j.fluid.2007.01.016 (2007).

24 Joback, K. G. & Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **57**, 233-243 (1987).

25 Wang, Q., Ma, P., Jia, Q. & Xia, S. Position group contribution method for the prediction of critical temperatures of organic compounds. *J. Chem. Eng. Data* **53**, 1103-1109 (2008).

26 Serat, F. Z., Benkouider, A. M., Yahiaoui, A. & Bagui, F. Nonlinear group contribution model for the prediction of flash points using normal boiling points. *Fluid Ph. Equilibr.* **449**, 52-59 (2017).

27 Jia, Q., Wang, Q. & Ma, P. Prediction of the Enthalpy of Vaporization of Organic Compounds at Their Normal Boiling Point with the Positional Distributive Contribution Method. *J. Chem. Eng. Data* **55**, 5614-5620, doi:10.1021/je1004824 (2010).

28 Gharagheizi, F., Ilani-Kashkouli, P., Acree Jr, W. E., Mohammadi, A. H. & Ramjugernath, D. A group contribution model for determining the vaporization enthalpy of organic compounds at the standard reference temperature of 298 K. *Fluid Ph. Equilibr.* **360**, 279-292 (2013).

29 Gharagheizi, F., Babaie, O. & Mazdeyasna, S. Prediction of vaporization enthalpy of pure compounds using a group contribution-based method. *Ind. Eng. Chem.* **50**, 6503-6507 (2011).

30 Li, R., Herreros, J. M., Tsolakis, A. & Yang, W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel* **304**, 121437 (2021).

31 Gharagheizi, F. Determination of normal boiling vaporization enthalpy using a new molecular-based model. *Fluid Ph. Equilibr.* **317**, 43-51, doi:https://doi.org/10.1016/j.fluid.2011.12.024 (2012).

32 Jia, Q., Yan, X., Lan, T., Yan, F. & Wang, Q. Norm indexes for predicting enthalpy of vaporization of organic compounds at the boiling point. *J. Mol. Liq.* **282**, 484-488, doi:https://doi.org/10.1016/j.molliq.2019.03.036 (2019).

33 Fissa, M. R., Lahiouel, Y., Khaouane, L. & Hanini, S. QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *J. Mol. Graph. Model.* **87**, 109-120 (2019).

34 Aouichaoui, A. R., Mansouri, S. S., Abildskov, J. & Sin, G. Uncertainty estimation in deep learning-based property models: Graph neural networks applied to the critical properties. *AIChE J.*, e17696 (2022).

35 Yao, X. *et al.* Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **62**, 217-225, doi:https://doi.org/10.1016/S0169-7439(02)00017-5 (2002).

36 Banchero, M. & Manna, L. Comparison between multi-linear-and radial-basis-function-neural-network-based QSPR models for the prediction of the critical temperature, critical pressure and acentric factor of organic compounds. *Molecules* **23**, 1379 (2018).

37 Egolf, L. M., Wessel, M. D. & Jurs, P. C. Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **34**, 947-956, doi:10.1021/ci00020a032 (1994).

38 Katritzky, A. R., Mu, L. & Karelson, M. Relationships of Critical Temperatures to Calculated Molecular Properties. *J. Chem. Inf. Comput. Sci.* **38**, 293-299, doi:10.1021/ci970071q (1998).

39 Turner, B. E., Costello, C. L. & Jurs, P. C. Prediction of Critical Temperatures and Pressures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **38**, 639-645, doi:10.1021/ci9800054 (1998).

40 Sola, D., Ferri, A., Banchero, M., Manna, L. & Sicardi, S. QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method. *Fluid Ph. Equilibr.* **263**, 33-42, doi:https://doi.org/10.1016/j.fluid.2007.09.022 (2008).

41 Sobati, M. A. & Abooali, D. Molecular based models for estimation of critical properties of pure refrigerants: Quantitative structure property relationship (QSPR) approach. *Thermochim. Acta* **602**, 53-62, doi:https://doi.org/10.1016/j.tca.2015.01.006 (2015).

42 Espinosa, G., Yaffe, D., Arenas, A., Cohen, Y. & Giralt, F. A Fuzzy ARTMAP-Based Quantitative Structure−Property Relationship (QSPR) for Predicting Physical Properties of Organic Compounds. *Ind. Eng. Chem.* **40**, 2757-2766, doi:10.1021/ie0008068 (2001).

43 Gharagheizi, F. & Mehrpooya, M. Prediction of some important physical properties of sulfur compounds using quantitative structure–properties relationships. *Mol. Divers.* **12**, 143-155 (2008).

44 Yao, X. *et al.* Radial basis function neural network based QSPR for the prediction of critical pressures of substituted benzenes. *Comput. Chem.* **26**, 159-169, doi:https://doi.org/10.1016/S0097-8485(01)00093-6 (2002).

45 Catoire, L. & Naudet, V. A unique equation to estimate flash points of selected pure liquids application to the correction of probably erroneous flash point values. *J. Phys. Chem. Ref. Data* **33**, 1083-1111 (2004).

46 Katritzky, A. R., Stoyanova-Slavova, I. B., Dobchev, D. A. & Karelson, M. QSPR modeling of flash points: An update. *J. Mol. Graph. Model.* **26**, 529-536 (2007).

47 Pan, Y., Jiang, J. & Wang, Z. Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network. *J. Hazard. Mater.* **147**, 424-430 (2007).

48 Carroll, F. A., Lin, C.-Y. & Quina, F. H. Improved prediction of hydrocarbon flash points from boiling point data. *Energy Fuels* **24**, 4854-4856 (2010).

49 Liu, X. & Liu, Z. Research Progress on Flash Point Prediction. *J. Chem. Eng. Data* **55**, 2943-2950, doi:10.1021/je1003143 (2010).

50 Carroll, F. A., Lin, C.-Y. & Quina, F. H. Simple method to evaluate and to predict flash points of organic compounds. *Ind. Eng. Chem.* **50**, 4796-4800 (2011).

51 Gharagheizi, F., Eslamimanesh, A., Mohammadi, A. H. & Richon, D. Empirical Method for Representing the Flash-Point Temperature of Pure Compounds. *Ind. Eng. Chem.* **50**, 5877-5880, doi:10.1021/ie102246v (2011).

52 Godinho, J. M., Lin, C.-Y., Carroll, F. A. & Quina, F. H. Group contribution method to predict boiling points and flash points of alkylbenzenes. *Energy Fuels* **25**, 4972-4976 (2011).

53 Saldana, D. A. *et al.* Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. *Energy Fuels* **25**, 3900-3908 (2011).

54 Mathieu, D. & Alaime, T. Insight into the contribution of individual functional groups to the flash point of organic compounds. *J. Hazard. Mater.* **267**, 169-174 (2014).

55 Phoon, L. Y., Mustaffa, A. A., Hashim, H. & Mat, R. A review of flash point prediction models for flammable liquid mixtures. *Ind. Eng. Chem.* **53**, 12553-12565 (2014).

56 Le, T. C., Ballard, M., Casey, P., Liu, M. S. & Winkler, D. A. Illuminating flash point: comprehensive prediction models. *Mol. Inform.* **34**, 18-27 (2015).

57 Sun, X. *et al.* Assessing Graph-based Deep Learning Models for Predicting Flash Point. *Mol. Inform.* **39**, 1900101 (2020).

58 Tetteh, J., Suzuki, T., Metcalfe, E. & Howells, S. Quantitative structure− property relationships for the estimation of boiling point and flash point using a radial basis function neural network. *J. Chem. Inf. Comput. Sci.* **39**, 491-507 (1999).

59 Dai, Y.-m. *et al.* Prediction of boiling points of organic compounds by QSPR tools. *J. Mol. Graph. Model.* **44**, 113-119 (2013).

60 Katritzky, A. R., Lobanov, V. S. & Karelson, M. Normal boiling points for organic compounds: correlation and prediction by a quantitative structure− property relationship. *J. Chem. Inf. Comput. Sci.* **38**, 28-41 (1998).

61 Jin, L. & Bai, P. QSPR study on normal boiling point of acyclic oxygen containing organic compounds by radial basis function artificial neural network. *Chemom. Intell. Lab. Syst.* **157**, 127-132 (2016).

62 Jin, L. & Bai, P. Prediction of the normal boiling point of oxygen containing organic compounds using quantitative structure–property relationship strategy. *Fluid Ph. Equilibr.* **427**, 194-201, doi:https://doi.org/10.1016/j.fluid.2016.07.015 (2016).

63 Osaghi, B. & Safa, F. QSPR study on the boiling points of aliphatic esters using the atom-type-based AI topological indices. *Rev. Roum. Chim.* **64**, 183-189 (2019).

64 Ericksen, D., Wilding, W. V., Oscarson, J. L. & Rowley, R. L. Use of the DIPPR database for development of QSPR correlations: Normal boiling point. *J. Chem. Eng. Data* **47**, 1293-1302 (2002).

65 Zhang, J. h., Liu, Z. m. & Liu, W. r. QSPR study for prediction of boiling points of 2475 organic compounds using stochastic gradient boosting. *J. Chemom.* **28**, 161-167 (2014).

66 Espinosa, G., Yaffe, D., Cohen, Y., Arenas, A. & Giralt, F. Neural network based quantitative structural property relations (QSPRs) for predicting boiling points of aliphatic hydrocarbons. *J. Chem. Inf. Comput. Sci.* **40**, 859-879 (2000).

67 Aldosari, M. N., Yalamanchi, K. K., Gao, X. & Sarathy, S. M. Predicting entropy and heat capacity of hydrocarbons using machine learning. *Energy and AI* **4**, 100054, doi:https://doi.org/10.1016/j.egyai.2021.100054 (2021).

68 Al Ibrahim, E. & Farooq, A. Octane Prediction from Infrared Spectroscopic Data. *Energy Fuels* **34**, 817-826, doi:10.1021/acs.energyfuels.9b02816 (2020).

69 Koenig, B. C., Ji, W. & Deng, S. Kinetic subspace investigation using neural network for uncertainty quantification in nonpremixed flamelets. *Proc. Combust. Inst.*, doi:https://doi.org/10.1016/j.proci.2022.07.226 (2022).

70 Veličković, P. *et al.* Graph attention networks. *arXiv:1710.10903* (2017).

71 Ryu, S., Lim, J., Hong, S. H. & Kim, W. Y. Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network. *arXiv:1805.10988* (2018).

72 Louis, S.-Y. *et al.* Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **22**, 18141-18148 (2020).

73 Zhu, W., Zhang, Y., Zhao, D., Xu, J. & Wang, L. HiGNN: A Hierarchical Informative Graph Neural Network for Molecular Property Prediction Equipped with Feature-Wise Attention. *J. Chem. Inf. Model.* **63**, 43-55, doi:10.1021/acs.jcim.2c01099 (2023).

74 Withnall, M., Lindelöf, E., Engkvist, O. & Chen, H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J. Cheminform.* **12**, 1, doi:10.1186/s13321-019-0407-y (2020).

75 Xiong, Z. *et al.* Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **63**, 8749-8760, doi:10.1021/acs.jmedchem.9b00959 (2020).

76 Ye, X.-b. *et al.* Molecular substructure graph attention network for molecular property identification in drug discovery. *Pattern Recognit.* **128**, 108659, doi:https://doi.org/10.1016/j.patcog.2022.108659 (2022).

77 Wiercioch, M. & Kirchmair, J. DNN-PP: A novel Deep Neural Network approach and its applicability in drug-related property prediction. *Expert Syst. Appl.* **213**, 119055, doi:https://doi.org/10.1016/j.eswa.2022.119055 (2023).

78 Omee, S. S. *et al.* Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns* **3**, 100491, doi:https://doi.org/10.1016/j.patter.2022.100491 (2022).

79 Aouichaoui, A. R. N., Fan, F., Mansouri, S. S., Abildskov, J. & Sin, G. Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models. *J. Chem. Inf. Model.*, doi:10.1021/acs.jcim.2c01091 (2023).

80 Jiang, D. *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.* **13**, 12, doi:10.1186/s13321-020-00479-8 (2021).

81 Kazakov, A. F., Muzny, C. D., Chirico, R. D., Diky, V. & Frenkel, M. D. NIST/TRC Web Thermo Tables-Professional Edition NIST Standard Reference Subscription Database 3, https://wtt-pro.nist.gov/wtt-pro/ (Accessed November 2020). (2002).

82 Wilding, W., Knotts, T., Giles, N. & Rowley, R. DIPPR® Data Compilation of Pure Chemical Properties. *Design Institute for Physical Properties, AIChE: New York, NY* (2020).

83 Wen, M., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858-1868, doi:10.1039/D0SC05251E (2021).

84 Thomas, M. *et al. Applications of artificial intelligence in drug design: opportunities and challenges. In: Artificial Intelligence in Drug Design*. 1-59 (2022).

85 Soleimany, A. P. *et al.* Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **7**, 1356-1367, doi:10.1021/acscentsci.1c00546 (2021).

86 Yang, C.-I. & Li, Y.-P. Explainable Uncertainty Quantifications for Deep Learning-Based Molecular Property Prediction. *ChemRxiv Preprint*, DOI: 10.26434/chemrxiv-22022-qt26449t (2022).

87 Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **148**, 241727 (2018).

88 Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **5**, 1963-1972 (2020).

89 Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365** (2019).

90 Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547-552, doi:10.1038/s41586-021-04184-w (2021).

91 Bhakat, S. Collective variable discovery in the age of machine learning: reality, hype and everything in between. *RSC Adv.* **12**, 25010-25024 (2022).

92 Vazquez-Salazar, L. I., Boittier, E. D., Unke, O. T. & Meuwly, M. Impact of the Characteristics of Quantum Chemical Databases on Machine Learning Prediction of Tautomerization Energies. *J. Chem. Theory Comput.* **17**, 4769-4785, doi:10.1021/acs.jctc.1c00363 (2021).

93 Ihme, M., Chung, W. T. & Mishra, A. A. Combustion machine learning: Principles, progress and prospects. *Prog. Energy Combust. Sci.* **91**, 101010, doi:https://doi.org/10.1016/j.pecs.2022.101010 (2022).

94 Sharma, P., Chung, W. T., Akoush, B. & Ihme, M. A Review of Physics-Informed Machine Learning in Fluid Mechanics. *Energies* **16**, 2343 (2023).

95 Huq, N. A. *et al.* Performance-advantaged ether diesel bioblendstock production by a priori design. *Proc. Natl. Acad. Sci.* **116**, 26421-26430 (2019).

96 Gaspar, D. J. *et al.* Top 13 Blendstocks Derived from Biomass for Mixing-Controlled Compression-Ignition (Diesel) Engines: Bioblendstocks with Potential for Decreased Emissions and Improved Operability. Pacific Northwest National Lab.(PNNL), 2021, Richland, WA (United States).

97 Lim, H. & Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **10**, 8306-8315 (2019).

98 Garai, J. Physical model for vaporization. *Fluid Ph. Equilibr.* **283**, 89-92, doi:https://doi.org/10.1016/j.fluid.2009.06.005 (2009).

99 Cho, J. *et al.* Bioderived ether design for low soot emission and high reactivity transport fuels. *Sustain. Energy Fuels* **6**, 3975-3988 (2022).

100 Xu, J. & Yu, C. Critical temperature criterion for selection of working fluids for subcritical pressure Organic Rankine cycles. *Energy* **74**, 719-733, doi:https://doi.org/10.1016/j.energy.2014.07.038 (2014).

101 Liu, B.-T., Chien, K.-H. & Wang, C.-C. Effect of working fluids on organic Rankine cycle for waste heat recovery. *Energy* **29**, 1207-1217, doi:https://doi.org/10.1016/j.energy.2004.01.004 (2004).

102 Zhang, T., Liu, L., Hao, J., Zhu, T. & Cui, G. Correlation analysis based multi-parameter optimization of the organic Rankine cycle for medium- and high-temperature waste heat recovery. *Appl. Therm. Eng.* **188**, 116626, doi:https://doi.org/10.1016/j.applthermaleng.2021.116626 (2021).

103 Zhang, X., Zhang, C., He, M. & Wang, J. Selection and Evaluation of Dry and Isentropic Organic Working Fluids Used in Organic Rankine Cycle Based on the Turning Point on Their Saturated Vapor Curves. *J. Therm. Sci.* **28**, 643-658, doi:10.1007/s11630-019-1149-x (2019).

104 Poling, B. E., Prausnitz, J. M. & O'Connell, J. P. *Properties of Gases and Liquids*. Fifth edn, (McGraw-Hill Education, 2001).

105 Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102-D1109 (2019).

106 Zhai, H., An, Q., Shi, L., Lemort, V. & Quoilin, S. Categorization and analysis of heat sources for organic Rankine cycle systems. *Renew. Sust. Energy Rev.* **64**, 790-805, doi:https://doi.org/10.1016/j.rser.2016.06.076 (2016).

107 Bell, I. H., Wronski, J., Quoilin, S. & Lemort, V. Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp. *Ind. Eng. Chem.* **53**, 2498-2508, doi:10.1021/ie4033999 (2014).

108 Aljundi, I. H. Effect of dry hydrocarbons and critical point temperature on the efficiencies of organic Rankine cycle. *Renew. Energy* **36**, 1196-1202, doi:https://doi.org/10.1016/j.renene.2010.09.022 (2011).

109 Zinsalo, J. M., Lamarche, L. & Raymond, J. Performance analysis and working fluid selection of an Organic Rankine Cycle Power Plant coupled to an Enhanced Geothermal System. *Energy* **245**, 123259, doi:https://doi.org/10.1016/j.energy.2022.123259 (2022).

110 ASTM International, ASTM D4814-21c. Standard Specification for Automotive Spark-Ignition Engine Fuel. (2021).

111 ASTM International, ASTM D975-22. Standard Specification for Diesel Fuel. (2022).

112 ASTM International, ASTM D1655-22. Standard Specification for Aviation Turbine Fuels. (2022).

113 Holladay, J., Abdullah, Z. & Heyne, J. Sustainable aviation fuel: Review of technical pathways. DOE/EE-2041. United States. DOI: 10.2172/1660415 (2020).

114 Van Rossum, G. Python Programming Language. *USENIX annual technical conference* **41**, 36 (2007).

115 Wang, M. *et al.* Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. (2019).

116 Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. *12th USENIX symposium on operating systems design and implementation*, 265-283 (2016).

117 Fioroni, G. M., Fouts, L., Christensen, E., Anderson, J. E. & McCormick, R. L. Measurement of Heat of Vaporization for Research Gasolines and Ethanol Blends by DSC/TGA. *Energy Fuels* **32**, 12607-12616, doi:10.1021/acs.energyfuels.8b03369 (2018).

118 Fioroni, G., Hays, C. K., Christensen, E. D. & McCormick, R. L. Reducing Sample Loss in Measurement of Heat of Vaporization of Ethanol/Gasoline Blends by Differential Scanning Calorimetry/Thermogravimetric Analysis. *SAE Int. J. Fuels Lubr.* **14**, 175-276 (2021).

119 Benzene, n-butyl-. Accessed October 8, 2021. https://webbook.nist.gov/cgi/cbook.cgi?ID=C104518&Mask=4#ref-12.

120 Steele, W. V., Chirico, R. D., Knipmeyer, S. E. & Nguyen, A. Vapor Pressure, Heat Capacity, and Density along the Saturation Line: Measurements for Benzenamine, Butylbenzene, sec-Butylbenzene, tert-Butylbenzene, 2,2-Dimethylbutanoic Acid, Tridecafluoroheptanoic Acid, 2-Butyl-2-ethyl-1,3-propanediol, 2,2,4-Trimethyl-1,3-pentanediol, and 1-Chloro-2-propanol. *J. Chem. Eng. Data* **47**, 648-666, doi:10.1021/je010083e (2002).

121 Luning Prak, D. J. *et al.* Determining the Thermal Properties of Military Jet Fuel JP-5 and Surrogate Mixtures Using Differential Scanning Calorimetry/Thermogravimetric Analysis and Differential Scanning Calorimetry Methods. *Energy Fuels* **34**, 4046-4054, doi:10.1021/acs.energyfuels.9b04028 (2020).