Towards Self-Consistent Graph Neural Networks for Predicting the Ideal Gas Heat Capacity, Enthalpy, and Entropy

Adem R. N. Aouichaoui,a Simon Müller,b Jens Abildskova\*

a Dept. of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads 228A, 2800 Kgs. Lyngby, Denmark

b Institute of Thermal Separation Processes, Hamburg University of Technology, Hamburg, Germany

ja@kt.dtu.dk

Abstract

Ideal gas heat capacity correlations are indispensable for modelling energy systems and evaluating process efficiency. While most correlations are empirical in nature, few are theoretically motivated, where the model parameters reflect physical quantities relating to the molecule. These however are rarely modelled through quantitative structure-property relationships, which hinders extending their applicability to new compounds. This work provides a realisation of a hybrid model that combines data-driven modelling in the form of a graph neural network that outputs a set of parameters used for the ideal gas heat capacity correlation. The study covered over 22,000 data points across 1,909 organic compounds resulting in a mean absolute error of 31.97 J/mol-K, a mean relative error of 11.63% and a correlation coefficient of 0.97.

**Keywords**: Graph neural networks, Hybrid modelling, Property prediction, QSPR

* 1. Introduction

*In-silico* evaluation of the thermophysical properties of molecules is an important prerequisite to performing large-scale screening of the chemical design space to identify suitable candidates for various applications. Flammability properties such as flash point and autoignition temperatures allow the identification of a chemical's safe processing and storage conditions. Critical point properties can be used in the cubic equation of states to perform P-V-T calculation and identify the process condition of chemicals. Quantitative structure-property relations (QSPRs) are mathematical models that relate the structural information of a given compound to the target property of interest. However, some properties are not only influenced by the structure of a molecule but also by some intensive variables such as temperature, pressure, and composition. Such properties include density, vapour pressure, thermal conductivity, and heat capacities (ideal gas and liquid). The latter plays an important role in modelling energy systems and enables performing energy balances, used in process simulation and technology evaluation.

* 1. Heat Capacity Correlation

The heat capacity of ideal gases () is defined as the amount of energy required to change the temperature of one mole of vapour by one degree assuming no intermolecular interactions. Most correlations developed are polynomial or exponential correlations as a function of the temperature, with a set of compound-specific constants determined through regression. In this work, we focus on the model developed by Aly and Lee (1981), which is derived from statistical mechanical formulae along with a series of assumptions and simplifications. This implies that the parameters used to develop the correlation have physical meaning, bringing the model closer to a true first-principle model. Based on a series of assumptions (“Born-Oppenheimer approximation for nuclear and electronic wave functions, the use of harmonic oscillator for vibrational motion and a decoupling of the vibrational and rotational motions with the grid rotator”), the total molecular energy can be separated into translational, rotational, and vibrational energies. Aly and Lee reduced the contribution to the overall to contributions from the most dominant vibrational characteristic temperature (), an electronic contribution () and a contribution from internal bond rotation () as seen in Eq.(1)

|  |  |
| --- | --- |
|  | (1) |

The contribution to the heat capacity from vibrational energy is expressed in Eq.(2),

|  |  |
| --- | --- |
|  | (2) |

Where B is a constant and C is a characteristic temperature. The contribution to the heat capacity from electronic energy is expressed in Eq.(3),

|  |  |
| --- | --- |
|  | (3) |

Where D and E are similar to Eq.(2) are a constant and characteristic temperature respectively. The contribution from the internal rotational bond is assumed constant for each compound as seen in Eq.(4).

|  |  |
| --- | --- |
|  | (4) |

Combining Eq.(2)-Eq.(4) yields the expression for seen in Eq.(5).

|  |  |
| --- | --- |
|  | (5) |

An analytical solution can readily be obtained to calculate the enthalpy and entropy of ideal gases through the fundamental relations relating these quantities to the heat capacity seen in Eq.(6). Integration constants are not of any concern as enthalpy and entropy are expressed as differences () w.r.t. a reference state.

|  |  |
| --- | --- |
|  | (6) |

Despite the integral role the property plays, QSPRs for predicting this quantity are few, especially for the ideal gases. Many still rely on the group-contribution (GC) method developed by Benson et al. (1969) and Joback and Reid (1987). Benson defined 250 groups and 40 corrections for nearest neighbors achieving 4-6 J/mol-K as an absolute error. Meanwhile, Joback and Ried defined 41 groups capable of describing organic compounds containing rings, sulfur, halogens, and nitrogen. The ideal gas heat capacity was modelled as a cubic polynomial in temperature with four coefficients (a, b, c, and d) which are modelled as a linear function of the group contributions. The study covered 288 compounds yielding a mean absolute percentage error of 1.59% and a mean absolute error of 5.9 J/mol-K. One advantage of developing QSPRs for this purpose is that solely the structural information is needed alongside the desired temperature for evaluation. However, one drawback of the GC methods is their linear nature, complex group identification procedure and inability to account for missing contributions. The recent surge in machine learning (ML) has renewed interest in developing QSPR capable of extracting relevant structural information and generating molecular descriptors from general molecular identifiers such as the SMILES (simplified molecular input line entry system).

* 1. Data-Driven Quantitative Structure-property Relation (QSPR)

Graph neural networks (GNNs) have emerged as viable candidates for ML models to predict a wide range of thermophysical properties with great accuracy (Aouichaoui et al., 2023b). These models take as input a graph representation and by staking convolutional layers (also known as message passing layers), the information embedded (also known as latent/hidden representation) in the nodes (representing atoms) and edges (representing bonds) is transferred internally (operations called message and update functions) and aggregated (also known as readout) to produce a vector representation of the molecule. This vector is used as an input to a multi-layer perceptron (MLP) to produce the target output. While GC methods can be seen as white-box models, GNNs are completely black-box with very little interpretable aspect. This might hinder their wider applicability as the model trend when extrapolating is not known. It has been shown that such models, despite having good performance metrics, might produce nonsensical results when extrapolating such as melting points being above boiling points (Aouichaoui et al., 2023b). This is largely due to the models not being constrained by physical laws. Recent trends in data-driven modelling have emphasized the need to add logic to govern the learnings of the model and to constrain it to generate meaningful and consistent outputs. Semi-parametric hybrid modelling combines data-driven and mechanistic modelling and has proven to generate accurate and physically meaningful outputs (Medina et al., 2022). In this work, we present one realization of hybrid modelling using GNNs for functional property prediction, in this case, the ideal gas heat capacity.

* 1. Methods
		1. Decoupled Serial Hybrid Modelling Framework

The hybrid modelling approach proposed herein uses five individual GNNs to predict the five parameters (A, B, C, D and E) needed for the correlation shown in Eq.(5). The framework “decouples” the initial identification of the five parameters, which are predetermined initially through least-square fitting using a range values at different temperatures. A schematic of the framework can be seen in Figure 1.

* + 1. Graph Neural Network Model: Attentive Fingerprint (AFP)

The attentive fingerprint (AFP) model by Xiong et al. (2020) uses the attention mechanism to enable the model to learn weights when aggregating messages across the molecule. This enables the model to focus on important substructures relevant to the target property. The attention mechanism consists of three steps: alignment (combines the hidden representation of a node with that of its neighbours), weighting (assigns a weight coefficient summing up to one through a softmax activation) and context (produces an update to the node). This mechanism is applied on a node level and graph level as an alternative approach to readout. The algorithm can be found in the work by Xiong et al. (2020). The molecular information chosen for featurizing the graph is commonly used in GNN modelling and is identical to those used by Aouichaoui et al. (2023).

|  |
| --- |
|  |
| Figure 1: Decoupled serial hybrid modelling framework for the ideal gas heat capacity |

* + 1. Data

The data used in this work are obtained from the Design Institute for Physical Properties under the American Institute of Chemical Engineers (AIChE-DIPPR) (Rowley et al., 2019). A total of 1,909 organic compounds containing halogens, nitrogen, and sulphur with at least five measurements (a total of 22,962 data points) and temperature ranges from 25K to 6000K were chosen. The data covered experimental (88 compounds) and predicted values, with the main prediction method being density function theory (DFT) calculations using B3LYP/6-311+G(3df,2p) frequency evaluation.

* + 1. Model development

Initially, the correlation parameters in Eq.(5) were determined through least squared error fitting using differential evolution (Storn and Price, 1997). A constraint was set that all parameters had to be positive since they represent characteristic temperatures. Then, five GNNs were trained independently using the same 90%-5%-5% randomly split compounds for training, validation and testing respectively. The models were trained for a maximum of 500 epochs and used early stopping to safeguard against overfitting. The model was trained using the ADAM optimizer with a scheduled learning rate similar to the study by Aouichaoui et al., (2023). The hyperparameters consisted of the number of node-level and graph-level message-passing layers as well as the length of the final representation produced. The number of MLP layers was fixed to 2 with the hidden layer being half of the length of the molecular feature produced. The hyperparameters were determined using a grid search where the number of layers ranged from one to four, while the length of the molecular feature vector could take the values 1024, 512, 256 or 128.

* 1. Results & Discussion

The preliminary fit of the parameters for Eq.(5) was successful yielding an overall coefficient of determination (R2) of 0.99, while the MAE and MRE were 0.93 J/mol-K and 0.43 % respectively. The lowest R2 was 0.93, while the maximum MAE and MRE were 10.20 J/mol-K and 4.5 % respectively. Table 1 shows the same metrics for predicting the five parameters through the GNNs. The model with the lowest validation MAE is chosen from the grid search.

Table 1: Performance metrics of the GNN on the parameters correlation: first value corresponds to that for training, the second for validation and the third for testing.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | A | B | C | D | E |
| R2 | 0.9260.9040.938 | 0.9550.9060.958 | 0.5230.3900.125 | 0.8890.4200.918 | 0.1060.0870.028 |
| MAE | 4.023.905.44 | 8.909.9111.75 | 219.30305.38336.342 | 10.1613.6911.97 | 317.82250.24344.08 |
| MRE | 24%27%28% | 12%14%12% | 18%26%26% | 22%23%18% | 35%37%36% |

Table 1 clearly shows that the GNN only succeeded in modelling the A, B and D parameters satisfactorily (the low validation R2 was due to an outlier). For the remaining parameters (C and E), the model fails to reproduce the results. This could indicate that there is no apparent relation between the structure and the determined C and E parameters, despite Eq.(5) being theoretically motivated. Therefore, we hypothesize that there are potentially multiple set of parameters that provides similar results due to collinearity and as such some of the determined values do not necessarily reflect the desired physical quantities. However, the endpoint of interest is as such, these were evaluated using the predicted parameters. Figure 2 shows a parity plot for the predictions from regressing the parameters (using differential evolution) and those obtained from predicting the parameters (using GNN).

|  |
| --- |
|  |
| Figure 2: Parity plot for calculated Cp\* through parameter fitting and the parameter prediction |

The results obtained showed surprisingly much better results than the individual parameter fitting. The hybrid model achieved an overall R2 of 0.97, an MAE of 31.97 J/mol-K and an MRE of 11.63%. The accuracy of the obtained stands in contrast with that of C and E, this can be attributed to i) deviations cancelling out, ii) the worst predicted parameters (C and E) are part of a quotient and thus their numerical value might not be affected by the errors or iii) the contribution from the quotients is negligible. Further insights can be obtained using sensitivity and collinearity analysis and will be the subject of future work. Compared to the GC methods, the results obtained cover a larger space, while achieving slightly lower performance metrics. This can be due to uneven data splitting (bias), the goodness of fit of the individual parameters or inconsistencies in the data. The obtained results were compared to a purely data-driven approach using GNN on a smaller subset of the data used herein (only experimental), which obtained an R2 of 0.99, MAE of 1.65 J/mol-K and MRE of 1.1% (Aouichaoui et al., 2023a). using the same subset of data, the current model achieves an R2, MAE and MRE of 0.96, 22.36 J/mol-K and 14.22% respectively. While the metrics favour the data-driven approach, it is important to note that the training methods were fundamentally different and as such it is not entirely possible to conclude which method is better.

* 1. Conclusions

The study presented a hybrid model for the ideal gas heat capacity that combines five GNNs to predict five compound-specific constants which are used in a theoretically motivated correlation. The results showed that the final prediction is good despite the presence of large errors on two of the five parameters. This can potentially be due to collinearity between the parameters. The presented model can predict the functional properties of various organic compounds solely based on the structural information and the state variable (in this case the temperature). This could potentially be extended to cover other similarly obtained properties such as densities and vapour pressure.

References

Aly, F.A., Lee, L.L., 1981. Self-consistent equations for calculating the ideal gas heat capacity, enthalpy, and entropy. Fluid Phase Equilibria 6, 169–179.

Aouichaoui, A.R.N., Cogliati, A., Abildskov, J., Sin, G., 2023a. S-GNN: State-Dependent Graph Neural Networks for Functional Molecular Properties, Computer Aided Chemical Engineering, 33 European Symposium on Computer Aided Process Engineering. Elsevier, pp. 575–581.

Aouichaoui, A.R.N., Fan, F., Abildskov, J., Sin, G., 2023b. Application of interpretable group-embedded graph neural networks for pure compound properties. Computers & Chemical Engineering 176, 108291.

Benson, S.W., Cruickshank, F.R., Golden, D.M., Haugen, G.R., O’Neal, H.E., Rodgers, A.S., Shaw, R., Walsh, R., 1969. Additivity rules for the estimation of thermochemical properties. Chem. Rev. 69, 279–324.

Joback, K.G., Reid, R.C., 1987. Estimation of pure component properties from group-contributions. Chemical Engineering Communications 57, 233–243.

Medina, E.I.S., Linke, S., Stoll, M., Sundmacher, K., 2022. Graph neural networks for the prediction of infinite dilution activity coefficients. Digital Discovery 1, 216–225.

Rowley, R.I., Oscarson, W.V., Giles, N.F., 2019. DIPPR data compilation of pure chemical properties.

Storn, R., Price, K., 1997. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. Journal of Global Optimization 11, 341–359.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., Zheng, M., 2020. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. J. Med. Chem. 63, 8749–8760.