Graph Neural Networks with Molecular Interaction Pooling for Pure-Component Vapor Pressure Prediction

Marco Hoffmann, Hans Hasse, and Fabian Jirasek

Laboratory of Engineering Thermodynamics, RPTU Kaiserslautern,
Erwin-Schroedinger-Strasse 44, 67663 Kaiserslautern

Abstract. We propose a novel hybrid approach based on graph neural networks (GNNs) and the physical Antoine equation that allows for consistently predicting the vapor pressure of pure components using only their molecular structure. For this purpose, we introduce a new pooling function, molecular interaction pooling, that, using attention, explicitly calculates the interaction between all atoms in the molecular graph, thereby adding expressivity to the model. Our model is superior in accuracy and applicability compared to state-of-the-art prediction methods.

1 Introduction

The vapor pressure p^s is a crucial property for designing and optimizing various processes in chemical engineering, e.g., absorption and distillation. For many standard components, experimental data on the vapor pressure are available, whose temperature dependence can be described with the Antoine equation

$$\ln\left(p^s/\mathrm{kPa}\right) = A - \frac{B}{C + T/\mathrm{K}}, \tag{1}$$

where A, B, and C are component-specific parameters that are fitted to p^s for the respective component. If these parameters have been fitted, p^s can usually consistently and reliably be predicted for a wide temperature range, exploiting the physical foundation of the Antoine equation. However, if no experimental data on the component of interest are available, these parameters can not be determined, making the entire prediction infeasible.

Therefore, methods to predict p^s for unstudied components are paramount. While several prediction methods have been put forward in the literature (see, e.g., Evangelista et al. [1]), they all come with significant drawbacks. They rely on static embeddings of the components, which are constructed through manual feature engineering and chemical intuition and may not be the most effective for the task. Furthermore, their use is limited to specific classes of molecules and/or they require information on additional properties of the components, which are not always available.

Graph neural networks (GNNs) represent a promising alternative to the established prediction methods, as they only require the information on the molecular structure, which is always available, and can derive a task-specific dynamic molecular embedding that is directly learned from the experimental data and is, therefore, much more expressive. Furthermore, they can efficiently be trained on large data sets.

In this work, we propose a new prediction method for p^s based on GNNs, which includes developing a novel *molecular interaction pooling function*. We combine the GNN with the physical Antoine equation to develop a powerful and robust hybrid model.

2 Data

The data used in this work were taken from the Dortmund Data Bank [2]. It comprises over 230,000 experimental data points on p^s of more than 25,000 components. The molecular graphs were constructed based on SMILES [3] strings using the rdkit package [4]. Hydrogen atoms were treated implicitly and added as node features, together with the information on the atom type, the number of bonds, the hybridization, whether the atom is aromatic, and whether the atom is part of a ring. The edge features contain the bond type, whether the bond is conjugated, part of a ring, or part of a stereoisomer. Categorical features were one-hot encoded.

3 Model

Our model consists of three main parts, which are described in detail below: The message passing layers to enrich the graph embedding by information exchange between the nodes; the molecular interaction pooling layer as read-out function to create a fixed-size embedding; the prediction head to compute Antoine parameters and then, using Eq. 1, the vapor pressure of the given component.

Message Passing. In the message passing phase, we use four layers of the graph attention network (GATv2) [5] with two attention heads per layer. From a chemical perspective, this step reflects how the atoms are influenced by their neighbors. Thus, after message passing, the node embedding contains not only information on the corresponding atom, but also on specific properties induced by their surrounding.

Molecular Interaction Pooling. The most intuitive pooling function for predicting vapor pressures is standard sum pooling, as the vapor pressure generally correlates with the molecular weight, and simply summing up all node embeddings can be interpreted as a 'count' of all heavy atoms in the molecule. However, sum pooling has two significant drawbacks: First, the individual node embeddings get diluted, and local structural information gets lost. Second, the model can not factor in interactions of atoms further apart than the message passing distance. For this reason, we propose a new pooling method called *molecular interaction pooling*. The idea is to calculate interactions between all nodes

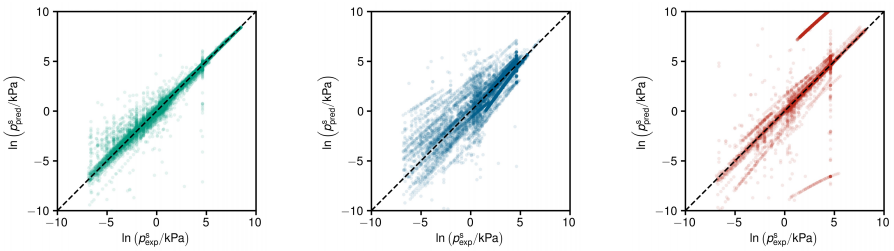
in the graph using a self-attention algorithm and then interpret the resulting context as a measure for interaction, summing these interactions to obtain a final fixed-size embedding. The molecular interaction pooling slightly outperforms standard sum pooling in our use case. We also tried the *set2set* pooling, which performed inferior to the other methods.

Prediction Head. The prediction head uses the fixed-size embedding to predict Antoine parameters (ref. Eq. 1) with a feed-forward neural network with three hidden layers and sigmoid functions to restrict the Antoine parameter ranges, such that the first derivative of the vapor pressure $\frac{\mathrm{d}p^\mathrm{s}}{\mathrm{d}T}$ is always positive ($B > 0$) and the first derivative of the enthalpy of vaporization $\frac{\mathrm{d}\Delta h_\mathrm{v}}{\mathrm{d}T}$ is always negative ($C < 0$). Finally, the logarithmic vapor pressure is calculated using the Antoine equation. During training, backpropagation adjusts the model parameters based on the deviations between the predicted and experimental data.

The data set was randomly split component-wise with 80 % of the components in the training set and 10 % in the validation and test set, respectively.

4 Results and Discussion

For benchmarking our novel method, we use two group-contribution methods that also only require molecular structure for prediction: SIMPOL.1 [6] and the method by Tu [7]. The results are shown in Fig. 1. It is evident that the



(a) Hybrid GNN (b) SIMPOL.1[6] (c) Tu[7]

Fig. 1: Parity plots of the predicted $\ln p^\mathrm{s}$ over the experimental test data for the hybrid GNN (proposed) and two benchmark methods. The GNN can predict all test data, SIMPOL.1 only 38 %, and the method by Tu only 27 %.

new model significantly outperforms both literature models as its predictions show much better agreement with the experimental test data. Moreover, both benchmark methods can only be applied to a subset of the components in the test set. The superiority of the proposed model is quantitatively underpinned by considering the median absolute percentage deviation (MAPE) on the test set. While the MAPE of the GNN is 6.33 %, those of SIMPOL.1 and the method by Tu are 57.27 % and 31.73 %, respectively.

We have also compared the scores of our new model with the reported scores of GNNs from the literature. Santana et al. recently presented the PUFFIN framework [8] and reported a mean squared error (here converted from decadic to natural logarithm) of 0.85 on their test set. As their data was not split by molecules but by temperature, PUFFIN does not need to extrapolate to new molecules on the test set. Despite this, our model shows a significantly better score of 0.30. The model of Lin et al. [9] was reported with a mean absolute error (calculated from their AARD score) of 0.48 on their test set. Our model achieves a mean absolute error of 0.20. Fig. 2 shows that our GNN even works well for complex multi-functional molecules for which p^{s} prediction is challenging.

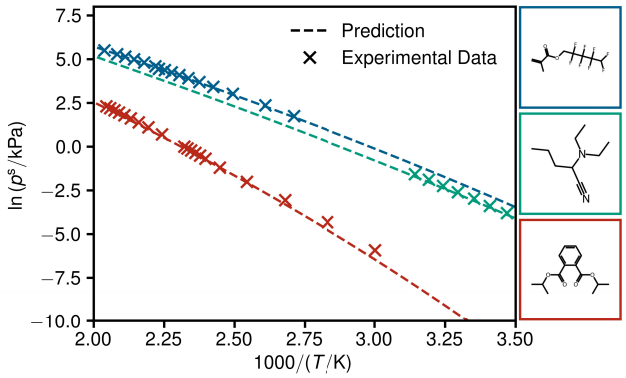


Fig. 2: Vapor pressure curves for three pure components from the test set as examples. The corresponding molecules for the vapor pressure curves are depicted on the right side. Experimental data are taken from the DDB [2].

5 Conclusion

In the present work, we introduce a novel model for predicting pure-component vapor pressures based only on the molecular structure. The model combines GATv2 [5] message passing layers with a newly developed attention-based molecular interaction pooling that produces a fixed-size embedding of the molecule. From this embedding, Antoine parameters are derived with a feed-forward neural network, whereby we enforce thermodynamic consistency. From the Antoine parameters, the vapor pressure can be obtained as a function of the temperature. The model significantly outperforms several benchmark methods and shows excellent results, making it a valuable tool for conceptual process design and optimization. In future work, we will extend the model, e.g., by incorporating data on the enthalpy of vaporization and the critical data into the training.

Acknowledgments. We gratefully acknowledge financial support by Carl Zeiss Foundation in the project "Process Engineering 4.0", as well as by Deutsche Forschungsgemeinschaft in the Priority Program 2363, and in the Emmy Noether Project of FJ.

References

1. Evangelista, N.S. et al.: Estimation of Vapor Pressures and Enthalpies of Vaporization of Biodiesel-Related Fatty Acid Alkyl Esters. Part 1. Evaluation of Group Contribution and Corresponding States Methods. Ind. Eng. Chem. Res. **56**, 2298–2309 (2017)
2. Dortmund Data Bank, 2024, www.ddbst.com
3. Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Science **28**, 31-36 (1988)
4. RDKit: Open-source cheminformatics. https://www.rdkit.org, version 2023.03.1
5. Brody, S. et al.: How Attentive are Graph Attention Networks?. arXiv pre-print, 2105.14491 (2021)
6. Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. Atmos. Chem. Phys. **8**(10), 2773–2796 (2008)
7. Tu, C.-H.: Group-contribution method for the estimation of vapor pressures. Fluid Phase Equilibr. **99**, 105–120 (1994)
8. Santana, V.V. et al: PUFFIN: A path-unifying feed-forward interfaced neural network for vapor pressure prediction. Chem. Eng. Sci. **286**, 119623 (2024)
9. Lin, Y.-H. et al: Advancing Vapor Pressure Prediction: A Machine Learning Approach with Directed Message Passing Neural Networks. ChemRxiv (2024)