Data Scarcity in Developing Property Prediction Models: Application of Multi-Task Transfer Learning

Adem R.N. Aouichaoui [C,S], Seyed Soheil Mansouri, Jens Abildskov and Gürkan Sin
Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, The technical university of Denmark, DTU, Lyngby, Kgs. Lyngby, Denmark
gsi@kt.dtu.dk

In absence of experimental data, the ability to predict various thermophysical properties related to chemicals of interest becomes crucial in many engineering applications. Some of them are essential in order to perform phase-equilibria calculations, energy balances and the evaluation of process alternatives [1]. One way to perform these predictions is by relating the molecular structure in a machine-readable format (molecular descriptors) to the property of interest. Such models are referred to as *"Quantitative Structure-Property Relationships"* or QSPR models [2]. Recent developments in the field of Deep Learning (DL) and especially Graph Neural Networks (GNN) have eliminated the need choosing a suitable molecular descriptor, as they are able to learn an optimal representation from a molecular graph and map them to the target property through backpropagation [3], [4]. Traditionally, models are built to predict one specific property or target. However, DL models are able to predict several properties simultaneously also known as multitask learning [5]. Here the models might improve their performance through inductive transfer learning i.e. while learning to predict property "A", the model might need less effort to learn how to predict property "B" [5] . This is especially relevant in cases where good quality experimental data are scarce [6].

In this work, we will demonstrate the model performance based on multi-task deep learning as compared to a series of disjoint models in predicting closely related properties e.g. flammability related properties covering the Lower Flammability Limit (LFL), Upper Flammability Limit (UFL) volumes and the Flash Point (FPT) and Auto-Ignition (AIT) temperatures.

References

[1] J. Frutiger, I. Bell, J. P. O'Connell, K. Kroenlein, J. Abildskov, and G. Sin, "Uncertainty assessment of equations of state with application to an organic Rankine cycle†," *Mol. Phys.*, vol. 115, no. 9–12, pp. 1225–1244, 2017.

[2] N. D. Austin, N. V. Sahinidis, and D. W. Trahan, "Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques," *Chem. Eng. Res. Des.*, vol. 116, pp. 2–26, 2016.

[3] C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, and K. F. Jensen, "Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction," *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 1757–1772, Aug. 2017.

[4] K. Yang *et al.*, "Analyzing Learned Molecular Representations for Property Prediction," *J. Chem. Inf. Model.*, vol. 59, no. 8, pp. 3370–3388, 2019.

[5] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv*, no. May, Jun. 2017.

[6] A. M. Schweidtmann, J. G. Rittig, A. König, M. Grohe, A. Mitsos, and M. Dahmen, "Graph Neural Networks for Prediction of Fuel Ignition Quality," *Energy & Fuels*, vol. 34, no. 9, pp. 11395–11407, Sep. 2020.