INSTITUT VERFAHRENSTECHNIK
Lehrstuhl Systemverfahrenstechnik
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg
www.ovgu.de/ivt/svt

Topic of Masters thesis
Graph Neural Networks for the prediction of environmental indicators of molecules

October 19, 2020

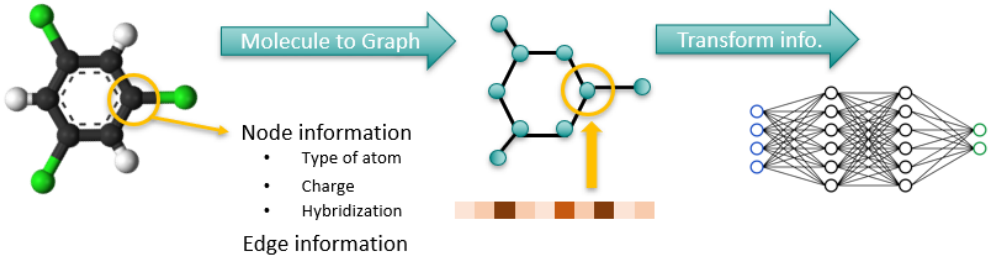
Motivation:

When moving towards a sustainable chemical industry, commonly used toxic substances have to be replaced with more environmentally benign options. In order to find sustainable replacements, one has to determine the environmental, healthy and safety conditions (EHS) of such molecules [1]. However, given the large space of possible molecules, the experimental determination of the EHS of all possible molecules turns to be an impossible task. Therefore, many models have been created to predict EHS based on the molecular structure solely. These models are known as quantitative structure-activity relationship (QSAR) models [2].

Practically all the QSAR methods require that the user pre-selects and calculates the appropriate molecular descriptors to serve as inputs to a regression or classification machine learning model, which is a difficult task given the large amount of possible descriptors. By contrast, if one can find an end-to-end model that relates the molecular structure to the property of interest, molecular descriptors would be no longer needed.

Problem definition:

Recently, Graph Neural Networks (GNNs) [3] have shown to be a promising tool to predict structure related quantities. The goal of this Masters thesis, is to construct and train different GNNs to predict different molecular sustainable indicators (e.g. bioconcentration factor, carcinogenicity). The molecules are represented as mathematical graphs, where the nodes are the atoms and the edges represent the chemical bonds. Both the nodes and the edges contain chemical information that, together with the connecting information of the graph itself, serve as the molecular descriptors to predict the indicator of interest. The candidate will use both our internal databases and gather extra ones from the literature, clean data, construct GNNs, train them and test them.



Requirements:

- **Motivation:** The most important requirement is to be highly motivated to apply Machine Learning ideas to solve Chemical Engineering problems. If you are motivated working hard is easier.

- **Python coding:** You need an intermediate knowledge of Python and being comfortable writing and finding code solutions for the task that needs to be accomplished. If you have some other experience in coding (e.g. Matlab) and are willing and highly motivated to learn Python is also fine.

- **Machine learning:** Knowledge of general concepts from Machine learning is expected (e.g. backpropagation, ANN, training, train/validation/test split, regression/classification)

- **Physical chemistry:** Bachelors-level physical chemistry knowledge is needed to understand how to embed real molecules information into mathematical graphs.

Start of work: November 2020.

Duration: 6 months.

Supervisors:

- M.Sc. Edgar Ivan Sanchez Medina / sanchez@mpi-magdeburg.mpg.de

- M.Sc. Steffen Linke / linke@mpi-magdeburg.mpg.de

References:

[1] Steffen Linke, Kevin McBride, Kai Sundmacher; “Systematic Green Solvent Selection for the Hydroformylation of Long-Chain Alkenes". ACS Sustainable Chemistry & Engineering 8.29, 10795-10811, 2020.

[2] Ghosh S., Kar S., Leszczynski J. (2020) Ecotoxicity Databases for QSAR Modeling. In: Roy K. (eds) Ecotoxicological QSARs. Methods in Pharmacology and Toxicology. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0150-1_29

[3] Battaglia, Peter W., et al. "Relational inductive biases, deep learning, and graph networks." arXiv preprint arXiv:1806.01261 (2018).