Predicting CO2 solubility in solvent mixtures using graph neural networks

Ulderico Di Caprioa, Min Wua, Emine Kayahanb, Florence Vermeirec, Tom Van Gervend, Peter Hellinckxe, Steffen Waldherrf , Mumin Enis Leblebicia,\*

aCenter for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, 3590 Diepenbeek, Belgium

bBiomimetics Group, Division of Mechatronics, Biostatistics and Sensors (MeBioS), Department of Biosystems, KU Leuven, Willem de Croylaan 42, 3001 Leuven, Belgium

cChemical Reactor Engineering and Safety (CREaS), Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, 3001 Leuven, Belgium

dProcess Engineering for Sustainable Systems, Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, 3001 Leuven, Belgium

eFaculty of Applied Engineering, University of Antwerp, Groenenborgerlaan 171, 2000 Antwerp, Belgium

fUniversity of Vienna, Molecular Systems Biology (MOSYS), Department of Functional and Evolutionary Ecology, Faculty of Life Sciences, 1030 Vienna, Austria

\*Corresponding author (muminenis.leblebici@kuleuven.be)

Abstract

The anthropogenic CO2 emissions reinforce the global warming phenomena. Post-combustion absorption is the most mature technology to reduce the CO2 footprint. In the last decade, many solvent blends have been developed and tested to improve the process efficiency. However, testing such solvent blends relying only on laboratory experiments is expensive and laborious. Computational screening represents a viable solution for this challenge. A reliable model to predict the absorption performance of CO2 in solvent blends is needed to achieve this task. Graph neural networks (GNNs) are a machine learning technique that correlates a graph structure with its properties through a learned molecular representation. They have shown significant capabilities in correlating the physic-chemical properties of a mixture with the molecular structures of its components. This work aims to build and analyse a model correlating the chemical composition of a mixture with its CO2 solubility using GNN. The model considers the molecular structure of the components in the mixture, their concentrations, and other process parameters (i.e., process temperature and CO2 concentration in the gas) as input variables. The model output is the CO2 molar fraction within the liquid at equilibrium. Several model configurations were tested, and the most performant one employed three graph attention layers with an embedding size of 16. The final model returned a mean absolute percentage error and a root mean squared error of 26.86% and 0.011 on the entire test set, and 22.34% and 0.011 on test data involving only new molecular structures.

**Keywords**: CO2 capture, Graph neural networks, QSPR relationship, Solvation.

* 1. Introduction

Global warming is one of the main challenges humanity is facing. It is enforced by anthropogenic CO2 emissions. Many technologies are available to reduce carbon footprint; the most mature is post-combustion capture (Bui et al., 2018). This solution employs chemical solvents to absorb the CO2 molecules in a flue-gas stream. The state-of-the-practice employs aqueous monoethanolamine (MEA) solutions at 30%wt. as the absorbent liquid. The chemistry of the solvent significantly affects the efficiency of the process; for this reason, many studies focus on the experimental investigation of novel solvent blends, investigating the CO2 absorption capacity of both organic and ionic liquids (Aghel et al., 2022). Solvent design is a promising way to improve the efficiency of such a process; however, the experimental investigation of the solvent mixtures is very time-consuming and laborious. Digital solvent screening represents a viable solution for this challenge. However, a reliable model to predict the absorption performance of CO2 in solvent blends is needed to achieve this task. The absorption properties of the solvent blend are connected to the chemical properties of the mixture components; therefore, correlating these two properties is the most promising way to achieve the modelling task. In amine blends and non-ionic solvents, absorption is typically driven both by chemical and physical mechanisms. Consequently, modelling such behaviour is challenging. In recent years, machine learning (ML) techniques have shown outstanding capabilities in predicting solubilities (Vermeire et al., 2021). They rely on experimental data of a given system to predict its behaviour and performance in given scenarios. More specifically, graph neural networks (GNNs) are an ML technique that correlates a graph structure with its properties through a learned molecular representation. Such a modelling technique has already been applied to predict the CO2 solubility in ionic liquids (Jian et al., 2022). Many studies are available in the literature correlating the molecular structure of non-ionic liquids with absorption properties using ML techniques, such as Orlov (2021) and Orlov (2022). However, to the best of our knowledge, the literature lacks studies considering mixtures between organic components and water and the process conditions (i.e., temperature and CO2 partial pressure in the flue gas) employing GNNs. This work aims to fill this gap by proposing a modelling methodology which applies GNN to predict the CO2 solubility in non-ionic liquids. The model considers a mixture of up to two organic components and their concentration in water at different process conditions.

* 1. Methodology

The data to train the network was obtained from literature and involved mixtures of organic liquids with up to two organic components. Overall, the dataset contained 1902 data points involving 25 unique organic molecules. Each data point contains the chemical structure of the organic components and their mass fraction, the mass fraction of water, the absorption temperature, the equilibrium CO2 partial pressure and the molar fraction of the dissolved CO2 within the absorbent liquid.

All the information was used as network input except the molar fraction of the dissolved CO2, which was used as model output. The information about the chemical structure was coded into molecular graphs through SMILES. The following properties characterised each node of the graph: 1) whether the atom is a hydrogen bond acceptor, 2) if the atom is in a ring and the dimension of the ring it is in, 3) the number of lonely pairs, 4) electronegativity, 5) atomic number and 6) hybridisation. All the node features were reported as continuous variables. The following properties characterised each edge between the nodes: 1) the bond type (i.e., single, double, aromatic), 2) whether the bond is in a conjugated system, 3) whether the bond is in a ring, and 4) the bond length. All the edge features were reported as continuous variables except for the bond type, which was reported as a one-hot encoding. In addition, each molecular graph was characterised using the physical properties of the molecules, namely: 1) total number of hydrogen bond acceptor sites on the molecule, 2) total number of hydrogen bond donors on the molecule, 3) total number of hydrogen within the molecule, 4) total number of carbon within the molecule, 5) total number of oxygen within the molecule, 6) total number of nitrogen within the molecule, 7) total number of sulphur within the molecule, 8) total number of rotatable bounds within the molecule, 9) molecule topological surface area, 10) molecular radius, 11) molar mass. The physical molecular information was included to support the information obtained from the molecular graph. This information could be learned from the molecular graph; however, this would require a more complex graph layer structure. Therefore, an explicity of physical molecular information was added to the network to improve the efficiency of the model training. The node, bounds and molecular properties were obtained using RDKit, a cheminformatics toolkit. In order to test the network performance, part of the dataset was held out from the training and used as the test set; it contained 566 data points. All the data points referred to 2-methylpiperizine (2MPZ) were used in the test set. Such a part of the test set comprised 42 data points. Additionally, to evaluate the performance under process conditions, random data points involving mixtures containing monoethanolamine (MEA), methyldiethanolamine (MDEA) and piperazine (PZ) were included in the test set. The training set was employed for the network hyperparameter identification and its training, while the test set was never used in the training; it was used only to evaluate the performance of the final model following the model hyperparameter identification and training step.

Figure 1 reports the structure of the employed network. The information from the molecules was extracted using graph convolutional layers. This work tested two convolutional layers: the one reported by Kipf and Welling (2017) and Veličković et al. (2018). Now on, the former is referred to as GCN and the latter as GAT. The training involved their implementations included in PyTorch Geometric 2.2.0 and employed the default settings. The information obtained from the convolutional layer was embedded in arrays using average and maximum pooling for each molecule. Such operators get the average and the maximum values over each matrix column returned by the convolutional layers. In addition, a further array was employed, including molecular information calculated via RDKit. The three arrays were concatenated, generating an array that is molecule-specific. After that, the molecule-specific arrays of the two organic components were multiplied by their mass fraction and summed. The obtained array was used as input to a multilayer perceptron (MLP) with the process conditions to calculate the dissolved CO2 amount. Prior to the usage in the MLP, the following scaling techniques were applied to the input: the molecular features were standardised, the temperature input was normalised in the range [0,1], and the CO2 pressure was firstly logarithmically scaled and then normalised in the range [0,1]. The MLP comprised two hidden layers, the first containing 20 nodes and the following 5 nodes. The layer had many nodes as the input, and the output layer contained only one output neuron. The internal activation functions were rectified linear units, while the output node employed a sigmoid function. The activation functions were found by manual grid search.

The number of graph layers, the type of graph layers and the embedding size were chosen using a grid search. The identification of the network hyperparameters was performed using a 10-fold cross-validation. Each fold contained mixtures with at least one molecule not included in the other folds; this way, the score on the test folder represents the model performance when dealing with unknown molecular structures, and the network selection is more reliable. A full-factorial investigation was performed involving two graph layers (i.e., GCN and GAT), three sizes of the molecular embedding (i.e., 16, 32 and 64) and three amount of graph layers (i.e., 1, 2 and 3). The network showing the lower median values of the mean squared error on the validation folds was selected. All the networks were trained using Adam optimiser for 500 epochs with a learning rate of 3·10-3.



Figure 1. Structure of the employed Graph Neural Network. In this figure: Graph Convolutional Layers (GCL), Average Pooling (AP), Mean Pooling (MP), Molecular Features (MF), Graph feature (GF), Organic representation (OR).

The network showing the best performance in the 10-fold cross-validation was trained on the entire training set, applying ensemble learning over 30 networks. Each network was trained on the entire train set with different initial guesses of the model weights. The results of the 30 networks were averaged and re-scaled to the original output values scale.

* 1. Results and discussion

Among the various networks, the most performant one on the 10-fold cross-validation employed a structure containing three graph attention layers per molecule, with a molecular embedding size of 16. Figure 2 reports the performance of this model, including the parity plots and the metrics on the test set. The root mean squared error (RMSE) and the mean absolute percentage error (MAPE) were employed.



Figure 2. Parity plot and metrics of the model on the test set. a) Model performance on the entire test set. b) Model performance on the test molecules not included in the train set. The abbreviations are as follows: 2-methylpiperizine (2MPZ), sulfolane (SULF), ethylene glycol (MEG), 1-methylimidazole (1MIMI), Dimethyl sulfoxide (DMSO), N-methyl-2-pyrrolidinone (NMP).

Figure 2a reports the performance of the model on the entire test set. In this case, the predictions are made both on molecules not included in the training set and on those included in the training set but at different process conditions. In this case, the model returns a MAPE value of 26.86% and a RMSE of 0.011. In addition, most of the points related to the molecules already included in the train set fell in the range ± 20%. This result highlights the prediction accuracy of the model over the process conditions.

Figure 2b reports the model accuracy on molecules it was not trained on. In this figure, only the points related to the 2MPZ are included. The model showed a MAPE of 22.34% and a RMSE of 0.011 in these conditions. From the figure, it is possible to assess how most of the points are within the error range of ± 20% when the organic blend is composed of only 2MPZ or 2MPZ-NMP. However, the predictions significantly diverge for all the other mixtures. The training data included points involving the other molecules within the organic part with 2MPZ; however, they included MEA or MDEA instead of 2MPZ. We hypothesise that this is related to the fact that both MEA and MDEA do not include cyclic structures in the molecule and have only one nitrogen atom in their structure; on the other hand, 2MPZ have a cyclic structure and includes two nitrogen atoms. However, further investigations are required to prove this hypothesis.

* 1. Conclusions

In this work, we obtained a model estimating the CO2 solubility within an organic mixture, given the molecular structure of the components and the process conditions. The model was obtained employing graph neural networks. The model generalisation capabilities were evaluated on unseen molecular structures and process conditions. In both cases, the trained model showed acceptable prediction accuracy with a mean absolute percentage error of 26.88% on the overall test set and 22.34% on the test set, including only unseen molecular structures.

A continuation of this work would be a more systematic investigation of further network structures to improve the model accuracy since some bias toward some input variables is reported. In addition, a further improvement would be assessing the estimation of other physical parameters affecting the capture process efficiencies.

References

B. Aghel et al., 2022, Review on CO2 capture by blended amine solutions, International Journal of Greenhouse Gas Control, 119, 103715

M. Bui et al., 2018, Carbon capture and storage (CCS): the way forward, Energy & Environmental Science, 11, 1062-1176

Y. Jian et al., 2022, Predicting CO2 Absorption in Ionic Liquids with Molecular Descriptors and Explainable Graph Neural Networks, ACS Sustainable Chemistry & Engineering, 10, 50, 16681–16691

T.N. Kipf and M. Welling, 2017, Semi-Supervised Classification with Graph Convolutional Networks, arXiv, 1609.02907

A.A. Orlov et al., 2021, Chemoinformatics-Driven Design of New Physical Solvents for Selective CO2 Absorption, Environmental Science & Technology, 55, 22, 15542–15553

A.A. Orlov et al., 2022, Computational screening methodology identifies effective solvents for CO2 capture, Communications Chemistry, 5, 37

P. Veličković et al., 2018, Graph Attention Networks, arXiv, 1710.10903

F.H. Vermeire and W.H. Green, 2021, Transfer learning for solvation free energies: From quantum chemistry to experiments, Chemical Engineering Journal, 418, 129307