# Predictive parametrization of thermodynamic models using Machine Learning

## Accessing PC-SAFT pure-component parameters with neural network ensembles

Jonas Habicht, Christoph Brandenbusch, Gabriele Sadowski

*Simulation of chemical or biotechnological processes is a key concept to efficiently design and optimize industrial processes or production plants. Thereby, thermodynamic phase equilibria of complex systems have to be available using advanced models, such as PC-SAFT. PC-SAFT requires the parametrization of each component, which is usually performed by fitting to experimental data. Acquiring experimental data is a time-consuming and costly procedure, for some components even unfeasible. In this work we used neural networks (NNs) to predict PC-SAFT pure-component parameters without the need of experimental data.*

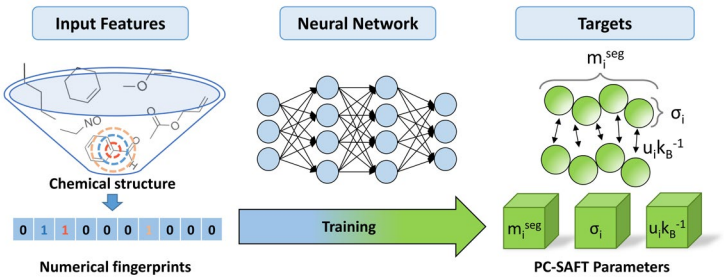


**Figure 1:** ML-framework to predict PC-SAFT pure-component parameters based on numerical fingerprints.

Classical neural networks are designed to map any arbitrary input features to targets (Figure 1). These targets can be for instance facial recognition, recognition of handwriting, or in our case the prediction of PC-SAFT pure-component parameters (three parameters for a non-associating molecule). NNs are trained using a training set of available data and comparing the true targets with the current NN-outputs. This allows the NN to "learn" underlying patterns in the training data. Having the trained NN in hand, the targets can now be predicted for an unknown molecule without any additional (experimental) input. To increase the statistical validity, multiple NNs are trained on slightly different training sets, creating a so-called NN-ensemble.

Defining suitable input features is crucial, as they must be processible by the NN and per definition non-experimental. The chemical structure of a molecule serves as non-experimental starting point to develop input features for the NN. Unfortunately, the structure itself is not processible by a NN and needs a transformation to a numerical representation.

In this work, a numerical fingerprint has been used as representation of the chemical structure. Starting from a central atom, substructures are derived by iteratively increasing the radius of neighboring atoms to collect larger substructures. This procedure is repeated for every atom of a molecule to collect more substructures, which are all stored in the numerical fingerprint.

The NN-ensemble was then trained on a dataset of ~300 molecules. The success of the training is visualized in Figure 2 (exemplary for one of the three parameters/targets) showing that predicted and validation data are in excellent agreement. With the trained NN-ensemble now in hand, PC-SAFT pure-component parameters for new molecules were predicted within seconds without the need to generate any experimental data, excellently describing experimental thermodynamic properties. Prospectively, NN-ensembles based on numerical fingerprints as input features can be used in an early state of process design to estimate thermodynamic properties/ phase behavior just from the chemical structure of a molecule.

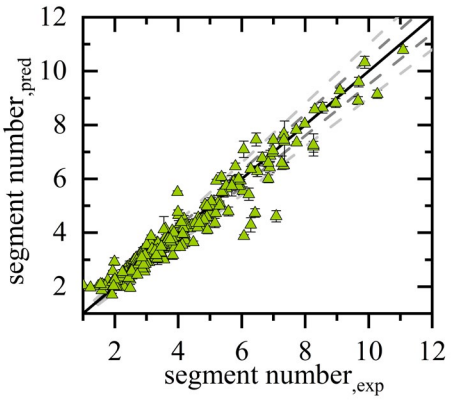


**Figure 2:** Parity plot for the predictions of the trained NN-ensemble for the segment number (target) compared to the original segment number.

**Contacts:**
jonas.habicht@tu-dortmund.de
christoph.brandenbusch@tu-dortmund.de
gabriele.sadowski@tu-dortmund.de

**Publications:**
J. Habicht, C. Brandenbusch, G. Sadowski, Fluid Phase Equilibria, 565, 113657 (2023)