SPT: A machine learning model for predicting phase equilibria

Benedikt Winter, Johanna Lindfeld, Luca Bosetti, André Bardow

Energy and Process System Engineering (ETH Zurich, 8092 Zurich, Switzerland)
bewinter@ethz.ch

The availability of property data is one of the major bottlenecks in the development of chemical and pharmaceutical processes, often requiring time-consuming and expensive experiments or limiting the design space to a small number of known molecules. This bottleneck of experimental data has been driving the continued development of predictive property models such as QSPR, group contribution methods, quantum-chemical, and molecular simulations. In recent times, machine learning has joined the more established property prediction models.

This contribution presents the SMILES-to-Properties transformer (SPT),[1,2] a natural language processing model that can predict pure component vapor pressures, melting temperatures, enthalpies of melting, and binary activity coefficients to a high degree of accuracy. These physical properties are required to calculate the phase equilibria most relevant to chemical and pharmaceutical separation processes: vapor-liquid (VLE), liquid-liquid (LLE), and solid-liquid equilibria (SLE).

As a natural language model, SPT first needs to understand the chemical grammar embedded in the SMILES code, which is a data-intensive step. To understand the grammar of SMILES, SPT is first pre-trained on millions of binary activity coefficients sampled from COSMO-RS. After the pretraining step, SPT is fine-tuned on available experimental data to reach high accuracy. This fine-tuning is not limited to the same physical property used during the pretraining but can be an unrelated one. We fine-tuned SPT on experimental data for activity coefficients,[3,4] vapor pressures,[3] melting points,[5] and enthalpies of melting.[5]

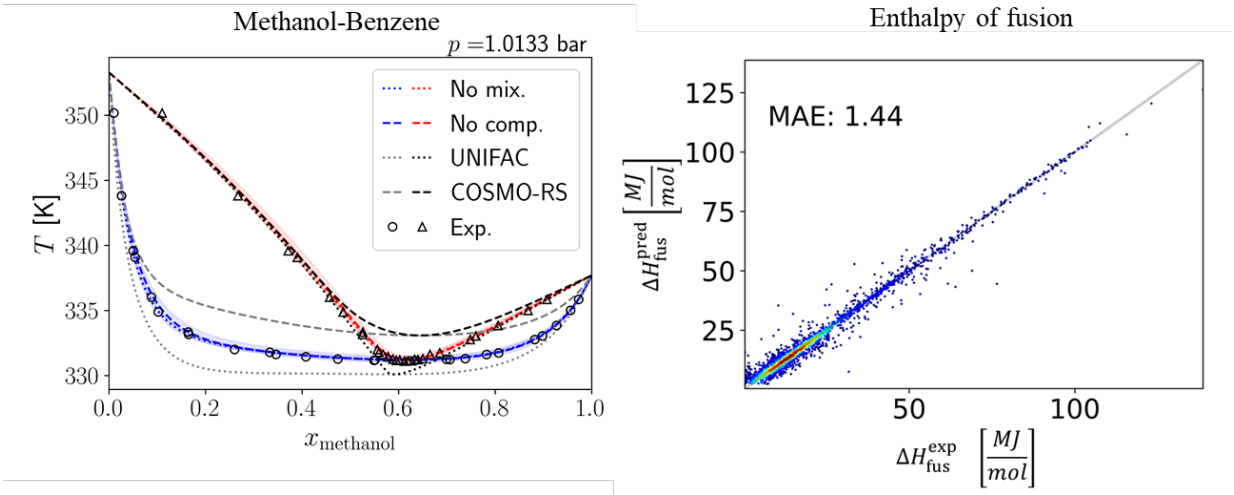

Figure 1: Left: VLE of the methanol/benzene mixture calculated using SPT, UNIFAC, and COSMO-RS. Right: parity plot of enthalpies of fusion predicted with SPT vs. experimental data using leave-one-out validation.

For the prediction of binary activity coefficients, SPT doubles the accuracy compared to commonly used models such as UNIFAC and COSMO-RS, even for mixtures where neither component is known. For the pure component properties, the melting temperature is predicted with a mean average error (MAE) of 17 K, and the enthalpies of melting with an MAE of 1.4 MJ/mol (See Fig. 1). Using the predicted vapor pressures to calculate the boiling temperature at 1.013 bar gives an average error of 2.5 K.

These high accuracies allow us to calculate phase equilibria of unknown components, thus giving the possibility to screen vast numbers of molecules *in-silico*. SPT can help with the separation of novel molecules, e.g., by predicting the properties of newly developed drugs and by enabling the optimization of existing processes through quick and extensive solvent screening. The predicted properties have been used to design separation processes involving unit operations that require the knowledge of phase equilibria, e.g., extraction (LLE), crystallization (LLE and SLE), distillation, and absorption (VLE).

In summary, SPT, as a novel tool to predict property data, has three major advantages: (i) SPT is very accurate compared to other predictive models (see figure 1), (ii) SPT can calculate properties of nearly arbitrary newly discovered molecules for quick process development and (iii) computations are almost instantaneous.

1) Winter, Benedikt; Winter, Clemens; Schilling, Johannes; Bardow, André (2022): A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing. In *Digital Discovery*. DOI: 10.1039/D2DD00058J.

2) Winter, Benedikt; Winter, Clemens; Esper, Timm; Schilling, Johannes; Bardow, André (2022): SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients. https://arxiv.org/abs/2209.04135

3) Dortmund Datenbank (2022). Available online at http://www.ddbst.com/, updated on 2022, checked on 2/6/2022.

4) Brouwer, Thomas; Schuur, Boelo (2019): Model Performances Evaluated for Infinite Dilution Activity Coefficients Prediction at 298.15 K. In *Ind. Eng. Chem. Res.* 58 (20), pp. 8903–8914. DOI: 10.1021/acs.iecr.9b00727.

5) Yaws, Carl L. (2006): The Yaws handbook of thermodynamic properties for hydrocarbons and chemicals. Houston, Tex.: Gulf; Lancaster: Gazelle Drake Academic.