# Hybrid Predictive Fluid Property Models – Integration of Physical Knowledge in Data-driven Matrix Completion Methods

*Fabian Jirasek[1], Nicolas Hayer[1], Thomas Specht[1], Julie Damay[2], Michael Bortz[2], Rima Abbas[3], Bastian Schmid[3], Hans Hasse[1]*

*[1]Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Germany*

*[2]Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany*

*[3]DDBST - Dortmund Data Bank Software & Separation Technology GmbH, Oldenburg, Germany*

Knowledge of thermodynamic properties of mixtures is of paramount importance in chemical engineering. Since experiments to determine thermodynamic property data are expensive and elaborate, prediction methods are required. While physical approaches such as UNIFAC and COSMO-RS have been successfully used for this purpose in the past decades, data-driven methods from machine learning (ML) open up new perspectives. In particular, matrix completion methods (MCMs) have yielded highly promising results for predicting properties of binary mixtures [1]. MCMs take advantage of the fact that data on binary mixtures can conveniently be stored in matrices, with rows and columns representing the components that make up the mixtures and the entries containing the mixture data. Due to the lack of experimental data, these matrices are usually very sparse. The prediction of the unobserved entries, i.e., of the properties of mixtures that have not been measured yet, constitutes a matrix completion problem, which is well known in ML and has become popular through the “Netflix Prize” [2], an open competition by Netflix that aimed at improving their recommender system for movies and TV shows.

While purely data-driven MCMs, trained only to the sparse matrices of mixture data, already outperform physical models in terms of predictive performance [1], hybrid approaches that incorporate physical knowledge, which is abundant in chemical engineering, promise unprecedented performance [3]. In the present work, we describe different approaches to hybridize MCMs for the prediction of thermodynamic properties, including the incorporation of physics-based prediction methods, physical pure component descriptors, and physical theory. The presented approaches are generic and can be applied for the prediction of any thermodynamic property of mixtures.

To illustrate the broad applicability as well as the different routes that can be taken for the hybridization, we discuss several examples, including a new hybrid model that combines an MCM with the UNIQUAC method and enables the prediction of activity coefficients in binary and multi-component mixtures as a function of temperature and composition. Furthermore, we discuss hybrid methods for the prediction of Henry's law constants and self-diffusion coefficients. Besides superior prediction accuracy compared to physics-based approaches, the presented methods are extremely convenient as they can be tailored very efficiently to specific needs (using, e.g., in-house data sets) or adapted when new experimental data become available.

[1] F. Jirasek, R.A.S. Alves, J. Damay, R.A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, H. Hasse: Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion, J. Phys. Chem. Lett. 11 (2020) 981-985.

[2] J. Bennett, S. Lanning: The Netflix Prize, Proceedings of KDD Cup and Workshop (2007).

[3] F. Jirasek, R. Bamler, S. Mandt: Hybridizing Physical and Data-driven Prediction Methods for Physicochemical Properties, Chem. Commun. 56 (2020) 12407.