Received: 20 August 2021 | Revised: 25 April 2022 | Accepted: 13 May 2022
DOI: 10.1002/aic.17753

AIChE JOURNAL

RESEARCH ARTICLE
Thermodynamics and Molecular-Scale Phenomena

Prediction of Henry's law constants by matrix completion

Nicolas Hayer | Fabian Jirasek | Hans Hasse

Laboratory of Engineering Thermodynamics (LTD), Technische Universität Kaiserslautern (TUK), Kaiserslautern, Germany

Correspondence
Fabian Jirasek, Laboratory of Engineering Thermodynamics (LTD), Technische Universität Kaiserslautern (TUK),
Erwin-Schrödinger Straße 44, Kaiserslautern 67663, Germany.
Email: fabian.jirasek@mv.uni-kl.de

Funding information
Bundesministerium für Wirtschaft und Energie; Carl-Zeiss-Stiftung

Abstract

Methods for predicting Henry's law constants H_{ij} are important as experimental data are scarce. We introduce a new machine learning approach for such predictions: matrix completion methods (MCMs) and demonstrate its applicability using a data base that contains experimental H_{ij} values for 101 solutes i and 247 solvents j at 298 K. Data on H_{ij} are only available for 2661 systems $i + j$. These H_{ij} are stored in a 101 × 247 matrix; the task of the MCM is to predict the missing entries. First, an entirely data-driven MCM is presented. Its predictive performance, evaluated using leave-one-out analysis, is similar to that of the Predictive Soave-Redlich-Kwong equation-of-state (PSRK-EoS), which, however, cannot be applied to all studied systems. Furthermore, a hybrid of MCM and PSRK-EoS is developed in a Bayesian framework, which yields an unprecedented performance for the prediction of H_{ij} of the studied data set.

KEYWORDS
gas solubility, Henry's law constant, machine learning, prediction, PSRK

1 | INTRODUCTION

Knowledge on the solubility of gases in solvents is essential for the design of many technical processes, such as gas absorption; and it is also needed for understanding many processes in nature. Gas solubility is usually described by Henry's law (cf., Equation S1), in which the key property is the Henry's law constant H_{ij}. The number of H_{ij} depends only on the temperature and the nature of the solute i and the solvent j. The solute is typically supercritical at the studied temperature, which is why it is called "gas." A large Henry's law constant H_{ij} corresponds to a low solubility and vice versa.

Experimental data on H_{ij} are scarce compared to the variety of possible combinations of relevant solutes and solvents. In the present work, we introduce new prediction methods for H_{ij} from the field of *machine learning* (ML): *matrix completion methods* (MCMs). Various types of MCMs have been proposed in the literature,[1–3] in particular for recommender systems,[4,5] and received a lot of attention through the Netflix Prize,[6] an open competition of Netflix aiming at improving their system for the prediction of user rating for movies and TV shows. In this work, we introduce MCMs for the prediction of H_{ij} at constant temperature in binary systems and thereby follow a Bayesian approach,[7–9] which is known to be robust to overfitting without requiring much parameter tuning.[10]

MCMs are highly interesting for predicting thermodynamic properties of binary systems. The idea behind this is that data for a given property of a binary system, such as the Henry's law constant H_{ij} of a solute i in a pure solvent j at a given temperature, can be stored conveniently in a matrix. The respective matrices containing the experimental data are typically very sparse, since the measurement of fluid properties is in general tedious and expensive and, in addition, the number of components and systems of interest is large. The prediction of the missing entries in such a matrix constitutes a matrix completion problem. We have recently introduced MCMs for the prediction of activity coefficients at infinite dilution in binary systems at constant temperature,[7,8] in which we give an in-depth discussion of the basic idea of applying MCMs for the prediction of thermodynamic mixture data. Here, we extend this approach to the prediction of H_{ij}.

In the present work, only pure solvents are considered and the temperature is fixed to 298.15 ± 1 K (labeled as 298 K here, for simplicity), such that H_{ij} is fully specified by specifying the components *i* and *j*. The temperature dependence of the Henry's law constant is highly interesting, but was excluded from the present study, which is focused on introducing new methods for predicting H_{ij}. However, these methods can be extended to include the temperature dependence of properties once they are established for the isothermal case. We have shown a possible approach to implement such an extension in a recent work[11] for the prediction of activity coefficients at infinite dilution γ_{ij}^{∞}, where we have modeled the dependence of γ_{ij}^{∞} on the temperature *T* by exploiting the fact that it can be well described by $\ln \gamma_{ij}^{\infty}(T) = A_{ij} + B_{ij}/T$ with system-specific, but temperature-independent, parameters A_{ij} and B_{ij} in many cases.

The accurate measurement of H_{ij} requires an extrapolation to the limiting case $x_i \rightarrow 0$, for which a series of experiments is necessary that makes these studies tedious. Therefore, experimental data on H_{ij} are missing for many practically relevant systems. This is why methods for predicting the Henry's law constant are so interesting.

Nevertheless, there are only comparatively few methods for predicting Henry's law constants so far. Most of these methods relate the Henry's law constant to physical component descriptors, mostly phenomenological descriptors like critical properties,[12] molecular descriptors like molecular masses and polarizability,[13,14] or SMILES representations.[15] These *Quantitative Structure Property Relationships* (QSPR)[16] are often based on nonlinear approaches like artificial neural networks or support vector machines. In some cases, techniques from ML have been used for descriptor selection, such as the *Replacement Method*[17,18] and *Genetic Algorithm* techniques.[17,19] All of these methods are restricted to a special class of systems: they either only consider aqueous solutions[13–15,17,18,20] or can only be applied for the prediction of the Henry's law constant of a single solute in different ionic liquids.[12,19] Since the scope of these methods is very restricted, they are not considered further here.

In contrast, *group-contribution equations-of-state* (GC-EoS), from which the Henry's law constant can be determined by well-established routes,[21] have a wider applicability. In GC-EoS the EoS is typically combined with a mixing rule that is based on a model of the Gibbs excess energy (G^E). Using a group-contribution G^E-model, such as UNIFAC[22] or modified UNIFAC (Dortmund),[23] then results in a GC-EoS, of which several have been proposed in the literature.[21,24,25] The EoS used in these approaches are often simple cubic EoS for which the G^E-mixing rules are known to give good results for a large variety of systems.[21,26]

The group-contribution concept enables predictions for systems for which no data are available. The prerequisite for carrying out this calculation is, however, that the group interaction parameters of the G^E-model are available. Parameter matrices including typical supercritical solutes, as they are encountered in gas solubility problems, have been established for GC-EoS[27]; however, the parameter tables are still far from covering all cases of interest.

One particularly successful GC-EoS, which has also been implemented in commercial process simulators, is the *Predictive Soave-Redlich-Kwong* (PSRK) EoS.[27,28] The PSRK-EoS (simply named PSRK in the following for brevity) is a combination of the cubic Soave-Redlich-Kwong EoS[29] with a mixing rule based on the original UNIFAC model.[22] The parameter tables for PSRK include many supercritical compounds. Specifically, the current public parameter table of the PSRK model distinguishes 81 main groups and comprises fitted pair-interaction parameters for 956 combinations of them.[27] Based on the reported parameters, a large number of components and systems can be modeled, and the PSRK model was also demonstrated to yield reliable predictions for many different systems,[27,30] although its predictive accuracy decreases for highly asymmetric systems.[31] However, note that there is still a substantial number of missing pair-interaction parameters of the PSRK model, namely for 2284 combinations of the present main groups, that have not been reported yet, which hampers its applicability. PSRK is used here as physical reference model for assessing the performance of the novel prediction methods based on matrix completion that are developed in this work. Furthermore, the physics-based PSRK is used in the development of a novel hybrid prediction method by combining it with a data-driven ML method as described in detail in the following sections.

The Henry's law constant $H_{ij}(T)$ can in principle also be determined from the pure component vapor pressure $p_i^{\text{vap}}(T)$ and the activity coefficient at infinite dilution $\gamma_{ij}^{\infty}(T)$ using information on the Poynting correction as well as on the pure component saturated vapor fugacity coefficient; for details, see Equation S4. However, determining $H_{ij}(T)$ in this way implies that the solute *i* is subcritical at the temperature *T*. In this case, typically Raoult's law would be used to describe the equilibrium condition of the component *i*, rather than using Henry's law, so that the Henry's law constant is not needed at all. Nevertheless, a substantial part of the experimental literature data on $H_{ij}(T)$ refers to this case. These data were included in the present study, but we emphasize that the main area of application of $H_{ij}(T)$ is the description of the solubility of supercritical components *i*.

This article is organized as follows: we first describe the data base for H_{ij} that we have used. We than introduce two different MCMs for predicting H_{ij}, one that is completely data-driven and one that constitutes a hybrid of a data-driven MCM and the physics-based PSRK. Subsequently, we present and discuss the results.

2 | DATA BASE

The experimental data on Henry's law constants H_{ij} of solutes *i* in solvents *j* at 298 K used in the present work were taken from the Dortmund Data Bank (DDB).[32] 298 K was chosen since at this temperature, by far the most data points for H_{ij} are reported in the DDB, as shown in Figure S1. The raw data on H_{ij} were preprocessed as described in the following. Data points that were labeled to be of poor quality in the DDB were excluded. Furthermore, only solutes and solvents for which data for at least two different binary systems were available were considered, as this is a prerequisite for the application of the leave-one-out analysis as described in detail below. Finally, for those binary systems for which multiple data points in the temperature range of 298.15 ± 1 K were available, the arithmetic mean of H_{ij} was calculated and used. The resulting data set comprises $I = 101$ solutes and $J = 247$ solvents and can, hence, be represented in a $I \times J$ matrix, which is depicted in Figure 1; information on the considered solutes and solvents is summarized in Tables S1 and S2, respectively. This matrix has 24,947 elements, but only 2661 of these are occupied with experimental data, corresponding to 10.7%. In Figure 1, the systems for which experimental data are available are represented as colored entries with the color code indicating the corresponding numerical value of H_{ij}, whereas the systems for which no experimental data are available are represented as black entries. The natural logarithm of H_{ij}, i.e., $\ln H_{ij}$, is thereby used in Figure 1 and throughout this work for scaling purposes.

Only 16 of the 101 solutes are supercritical at the considered temperature. This is an extremely small number, considering the importance of gas solubility. In order to have a sufficiently large data base, we did non differentiate between sub- and supercritical solutes in the present work and simply operated on all available data in the DDB.

It is interesting that the entries in a single row in Figure 1 show a fairly uniform color, i.e., for a given solute, the numbers of H_{ij} are similar for most solvents. In contrast, for a given solvent, the numbers of H_{ij} vary strongly, depending on the solute it is combined with. Furthermore, the color code indicating the values of H_{ij} in Figure 1 reveals a strong correlation between the critical temperature of a solute and its solubility: for solutes with lower critical temperature, in general higher H_{ij} are observed and vice versa.

There are, however, a few apparent exceptions: most of the considered solutes are hydrophobic and therefore substantially poorer soluble in water (H_2O) and heavy water (D_2O) than in other solvents, cf., labeled columns in Figure 1. Furthermore, the solute sulfur trioxide (SO_3) exhibits rather high Henry's law constants (poor solubilities) despite a comparatively high critical temperature; however, as for SO_3 only data for two solvents are available, this finding should not be overly interpreted.

Twenty-nine of the components are present both as solute and solvent in the data set, cf., Tables S1 and S2. The corresponding solute/solvent combinations would be pure components and were therefore not considered in the present work; for a detailed discussion of these cases, we refer to the Supporting Information.

For subcritical solutes, Henry's law constants can also be calculated from the solute's vapor pressure and its activity coefficient at infinite dilution. In principle, this could have been used for augmenting the data base on H_{ij} here. This option was considered but discarded, firstly, as it would have further increased the already large fraction of data for subcritical solutes, and, secondly, as the corresponding calculation requires assumptions on the fugacity coefficient of the solute and the Poynting correction, which introduce additional errors.

3 | MATRIX COMPLETION METHODS

Two different MCMs were used in this work for predicting Henry's law constants H_{ij} for binary systems at 298 K. Both MCMs are based on a Bayesian approach,[7,8] which considers random variables drawn from a probability distribution instead of scalar parameters and which enables the incorporation of prior knowledge in a straightforward manner, as described in detail below. Both MCMs are collaborative-filtering methods[5,33] that do not incorporate any direct information on the *pure components*, such as physical component descriptors, but use only the available *mixture data* for the binary systems, from which they infer so-called *latent variables* (LVs) during the training.

In both MCMs, the natural logarithm of H_{ij} is modeled as a stochastic function of LVs:

$$\ln H_{ij}^{\mathrm{MCM}} = \boldsymbol{u}_i^{\mathrm{T}} \cdot \boldsymbol{v}_j + b_i^u + b_j^v, \qquad (1)$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are (column) vectors of length K, whereas b_i^u and b_j^v are scalars. $\boldsymbol{u}_i$ and b_i^u represent the LVs of the solute i, $\boldsymbol{v}_j$ and b_j^v those of the solvent j. Hence, in both MCMs, each solute and each solvent is described by $K+1$ component-specific LVs, which are determined from data on the mixture property $\ln H_{ij}$ (all LVs are initially unknown

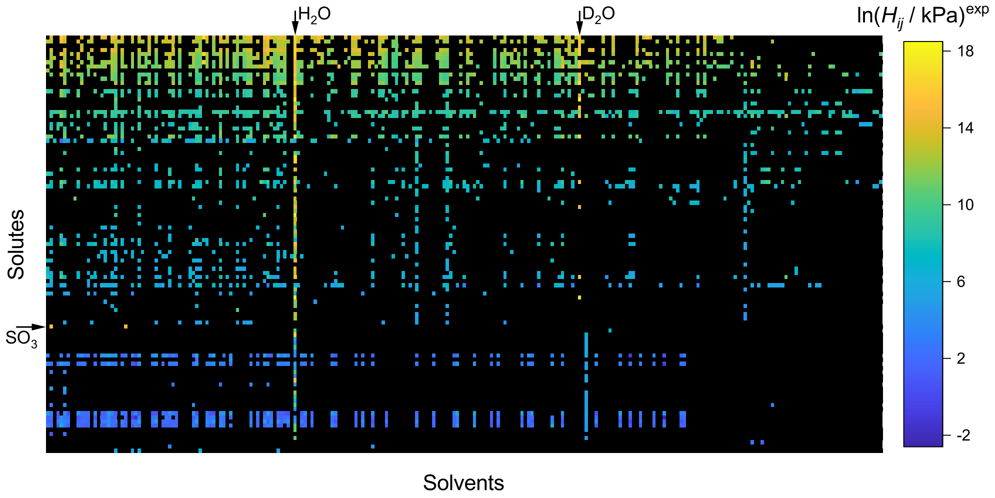

FIGURE 1 Matrix representing the experimental data on Henry's law constants H_{ij} of solutes i in pure solvents j at 298.15 ± 1 K as reported in the DDB[32] after preprocessing (see text). The color code indicates the numerical value of $\ln H_{ij}$. The order of the solvents is arbitrary, while the solutes are arranged according to their critical temperature T_c according to the DDB from low (top) to high (bottom).

and inferred from the training data on $\ln H_{ij}$ during the training of the MCMs). K is a hyperparameter of the models and was set to $K = 4$ in all cases based on preliminary studies using cross-validation; however, the presented MCMs are robust regarding variations of K as demonstrated in Figure S10.

The product $\boldsymbol{u}_i^{\mathrm{T}} \cdot \boldsymbol{v}_j$ in Equation (1) describes the contribution of specific pairwise interactions between solute i and solvent j to $\ln H_{ij}$, whereas b_i^u and b_j^v can be interpreted as a *general solubility* of a solute i and a *general dissolving capacity* of a solvent j, respectively, irrespective of specific binary interactions. In the following, we refer to b_i^u and b_j^v as *solute bias* and *solvent bias*, respectively, or summarize both under the term *component biases*. Such biases are also commonly considered for users and movies in recommender systems of, for example, movie streaming services, where they take into account that some users are generally more critical than others when rating movies, and that some movies are in general rated higher than others.[34] They turn out to improve the model also in the present application for predicting H_{ij}. The consideration of the solute bias b_i^u is motivated in particular by the observation that some solutes show poor solubility in almost all studied solvents whereas other solutes are highly soluble in most solvents, cf., Figure 1. A similar behavior was not observed for activity coefficients at infinite dilution, which we studied in our previous work.[7,8]

In the Bayesian approach that is used here, all data and LVs are modeled as random variables such that the MCMs are probabilistic models, that are defined by two probability distributions: *prior* and *likelihood*. The prior constitutes a probability distribution over the parameters of a model (LVs of the MCM here) *prior* to fitting the model to the training data. Hence, the prior contains a priori information on the LVs before the actual training step. The likelihood, on the other hand, describes the link between the training data and the LVs. The likelihood constitutes a probability distribution over the observable quantity ($\ln H_{ij}$ here) conditioned on the LVs, i.e., it specifies how the LVs manifest themselves in the data for $\ln H_{ij}$. The aim of Bayesian inference is finding the so-called *posterior*, which is the probability distribution over the LVs that is consistent with both the a priori information on the LVs (through the prior) and the evidence from the training data (through the likelihood). As inference method, variational inference[35,36] was chosen in this work.

From the posterior, i.e., the inferred LVs, $\ln H_{ij}$ can also be predicted for previously unreported binary systems following Equation (1). In each case, a probability distribution for $\ln H_{ij}$ is thereby predicted, which also provides information on the model uncertainties. In the following sections, the characteristics of the two MCMs developed in this work are discussed in more detail.

3.1 | Data-driven MCM

The first MCM is purely data-driven: its LVs are trained only to the sparse available experimental data for $\ln H_{ij}$ from the DDB, cf., Figure 1; no other information is used. We refer to this method as *MCM-data* in the following. Figure 2 shows an overview of how MCM-data is trained and used to predict $\ln H_{ij}$.

In the case of MCM-data, no information about the LVs is available prior to the training. Therefore, a rather broad, thus uninformative, probability distribution was used as prior here. Specifically, a normal distribution centered around zero and standard deviation $\sigma_{\mathrm{P}} = 1$ for $\boldsymbol{u}_i^{\mathrm{T}}$ and $\boldsymbol{v}_j$, and $\sigma_{\mathrm{P,CB}} = 10$ for b_i^u and b_j^v was chosen:

$$p(u_{i,k}) = \mathcal{N}(0, \sigma_{\mathrm{P}}), \quad \text{for } k = 1 \ldots K \tag{2}$$

$$p(v_{j,k}) = \mathcal{N}(0, \sigma_{\mathrm{P}}), \quad \text{for } k = 1 \ldots K \tag{3}$$

$$p(b_i^u) = \mathcal{N}(0, \sigma_{\mathrm{P,CB}}) \tag{4}$$

$$p(b_j^v) = \mathcal{N}(0, \sigma_{\mathrm{P,CB}}). \tag{5}$$

In general, the smaller the values for the standard deviations (σ_{P} and $\sigma_{\mathrm{P,CB}}$) are chosen, the stronger the LVs are restricted and the smaller is the influence of the training data on the LVs. In contrast, a very broad prior distribution (large values of σ_{P} and $\sigma_{\mathrm{P,CB}}$) barely constrains the LVs, such that the posterior is predominantly determined by the experimental data. The influence of the choice of the hyperparameters was investigated in preliminary studies. The values reported here represent good compromises between the extremes "too narrow, i.e., too restrictive" and "too broad, i.e., too irrelevant." However, the window, in which good results are obtained is wide and similar results as the ones presented below can also be obtained with other choices of the hyperparameters.

As likelihood, which models the probability of the data on $\ln H_{ij}$ conditioned to the LVs, a normal distribution with standard deviation $\sigma_{\mathrm{L}} = 0.2$ centered around $\boldsymbol{u}_i^{\mathrm{T}} \cdot \boldsymbol{v}_j + b_i^u + b_j^v$ was chosen:

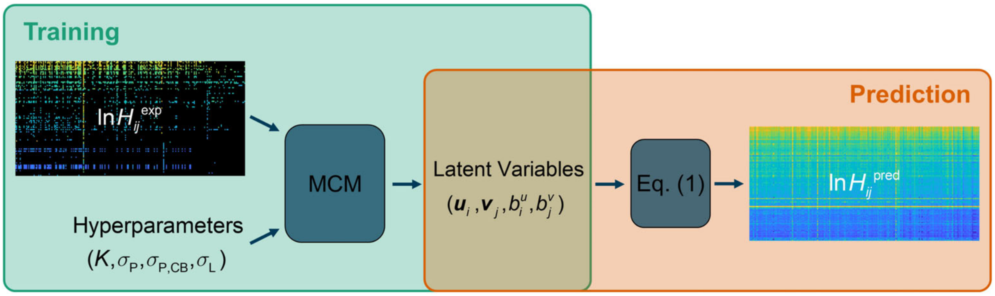
Training
$\ln H_{ij}^{\mathrm{exp}}$
Hyperparameters
$(K, \sigma_{\mathrm{P}}, \sigma_{\mathrm{P,CB}}, \sigma_{\mathrm{L}})$
MCM
Latent Variables
$(\boldsymbol{u}_i, \boldsymbol{v}_j, b_i^u, b_j^v)$
Prediction
Eq. (1)
$\ln H_{ij}^{\mathrm{pred}}$

FIGURE 2 Schematic illustration of the prediction of $\ln H_{ij}$ with MCM-data. The MCM is trained to experimental data on $\ln H_{ij}$ (exp) with specified hyperparameters. The inferred LVs are subsequently used with Equation (1) to obtain predictions (pred) for all possible solute-solvent combinations.

$$p\left(\ln H_{ij}|\boldsymbol{u}_i,\boldsymbol{v}_j,b_i^u,b_j^v\right)=\mathcal{N}\left(\boldsymbol{u}_i^{\mathrm{T}}\cdot\boldsymbol{v}_j+b_i^u+b_j^v,\sigma_{\mathrm{L}}\right)$$
$$=\mathcal{N}\left(u_{i,1}\cdot v_{j,1}+\cdots+u_{i,K}\cdot v_{j,K}+b_i^u+b_j^v,\sigma_{\mathrm{L}}\right), \tag{6}$$

The choice of the hyperparameter $\sigma_{\mathrm{L}}=0.2$ was motivated by the uncertainty of the available experimental data, i.e., twice the value found on average for the experimental uncertainty was chosen. All data points were thereby treated equally, i.e., $\sigma_{\mathrm{L}}=0.2$ was used throughout.

3.2 | HYBRID MCM

The second MCM is hybrid, as it is not only trained to (sparse) experimental data on $\ln H_{ij}$, but also incorporates information from the Predictive Soave-Redlich-Kwong (PSRK) equation-of-state[27] in the form of PSRK predictions. Consequently, we refer to this MCM as *MCM-hybrid* in the following. MCM-hybrid is based on the so-called *whisky approach* proposed for the prediction of activity coefficients at infinite dilution in previous work of our group[8] and therefore only briefly discussed here; we refer to Reference [8] for more details. Figure 3 shows an overview of how MCM-hybrid is trained and used to predict $\ln H_{ij}$.

As MCM-data, MCM-hybrid models $\ln H_{ij}$ according to Equation (1). However, in contrast to MCM-data, MCM-hybrid takes full advantage of the Bayesian approach to matrix completion by using an *informative* prior. The training of MCM-hybrid consists of two steps. In the first step, MCM-hybrid was trained to *simulated* data for $\ln H_{ij}$ that was generated with PSRK. With PSRK and its present public parameterization,[27] predictions for 7760 (31.1%) of all possible binary systems of the considered solutes and solvents can be obtained; hence, the matrix with this simulated data for the first training step is more densely occupied than the matrix with the experimental data, cf., Figure 1. During the first training step, the MCM infers (provisional) LVs of the solutes and solvents from the predictions of PSRK, cf., Equation (1). This step can be considered as *extracting* the physical knowledge on the solutes and solvents that is implicitly encoded in PSRK and explicitly provided in the form of PSRK predictions for $\ln H_{ij}$, and storing this knowledge in LVs. However, as the PSRK predictions are less reliable than the experimental data, the LVs obtained in this *pretraining* step are only preliminary and are therefore not directly used for predicting $\ln H_{ij}$. Instead, they are used to generate an *informative* prior for a second training step of the MCM. In the second training step, MCM-hybrid is, similarly to MCM-data, trained to the sparse set of available experimental data on $\ln H_{ij}$. The second step can be understood as a revision of the preliminary LVs (inferred from the PSRK predictions alone) based on the experimental data; we refer to this step as *refinement* step in the following. The refinement step of MCM-hybrid yields the final set of LVs that contain information from the PSRK predictions *and* the experimental data. In the pretraining step of MCM-hybrid, the same broad normal distribution as in MCM-data, i.e., a normal distribution centered around zero with standard deviation $\sigma_{\mathrm{P,CB}}=10$ for the component biases and $\sigma_{\mathrm{P}}=1$ for the remaining LVs, was used as prior:

$$p(u_{i,k})=\mathcal{N}(0,\sigma_{\mathrm{P}}),\quad \text{for}\,k=1\ldots K \tag{7}$$

$$p(v_{j,k})=\mathcal{N}(0,\sigma_{\mathrm{P}}),\quad \text{for}\,k=1\ldots K \tag{8}$$

$$p(b_i^u)=\mathcal{N}(0,\sigma_{\mathrm{P,CB}}) \tag{9}$$

$$p(b_j^v)=\mathcal{N}(0,\sigma_{\mathrm{P,CB}}). \tag{10}$$

A Cauchy distribution with scale $\lambda_{\mathrm{L}}=0.2$ was chosen as likelihood, which is in contrast to the training of MCM-data:

$$p\left(\ln H_{ij}|\boldsymbol{u}_i,\boldsymbol{v}_j,b_i^u,b_j^v\right)=\mathrm{Cauchy}\left(\boldsymbol{u}_i^{\mathrm{T}}\cdot\boldsymbol{v}_j+b_i^u+b_j^v,\lambda_{\mathrm{L}}\right)$$
$$=\mathrm{Cauchy}\left(u_{i,1}\cdot v_{j,1}+\cdots+u_{i,K}\cdot v_{j,K}+b_i^u+b_j^v,\lambda_{\mathrm{L}}\right). \tag{11}$$

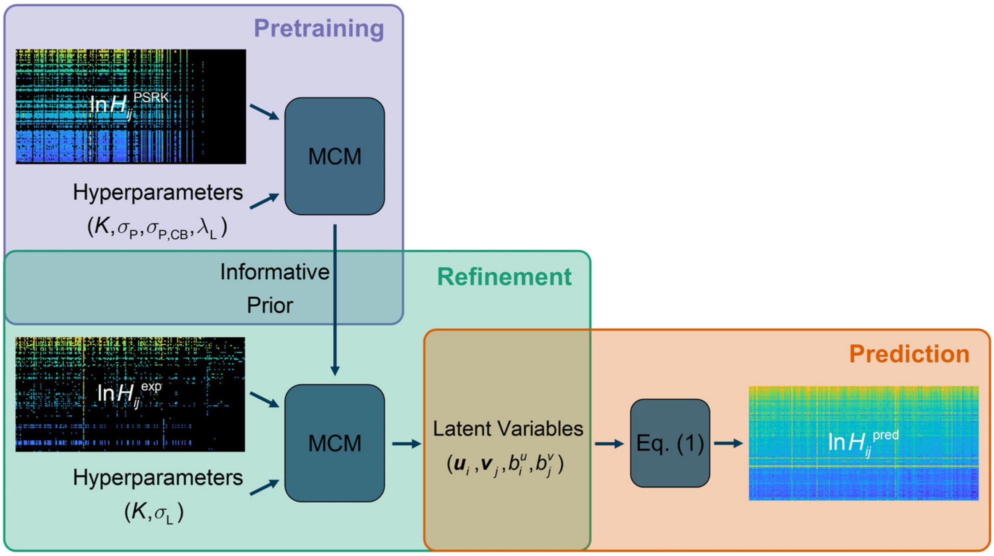
Pretraining
$\ln H_{ij}^{\mathrm{PSRK}}$
MCM
Hyperparameters
$(K,\sigma_{\mathrm{P}},\sigma_{\mathrm{P,CB}},\lambda_{\mathrm{L}})$
Informative Prior
Refinement
$\ln H_{ij}^{\mathrm{exp}}$
MCM
Hyperparameters
(K,σ_{L})
Prediction
Latent Variables
$(\boldsymbol{u}_i,\boldsymbol{v}_j,b_i^u,b_j^v)$
Eq. (1)
$\ln H_{ij}^{\mathrm{pred}}$

FIGURE 3 Schematic illustration of the prediction of $\ln H_{ij}$ with MCM-hybrid. In the pretraining step, the hyperparameters are specified and the MCM is trained to simulated data for $\ln H_{ij}$ from PSRK. The inferred (preliminary) LVs are used to generate an informative prior for the refinement step, in which the MCM is trained to experimental data on $\ln H_{ij}$ (exp). The resulting (final) LVs are subsequently used with Equation (1) to obtain predictions (pred) for all possible solute-solvent combinations.

The reason for using a Cauchy likelihood is that for some combinations of solutes and solvents, PSRK gives extremely (and unreasonably) large/small predictions for $\ln H_{ij}$ as shown in Figure S5. We attribute these extreme outliers to badly chosen binary interaction parameters of PSRK; the problematic predictions are basically limited to hydrochloric acid (HCl) dissolved in alcohols. To prevent a negative impact due to these obvious outliers in the pretraining step of MCM-hybrid, the Cauchy distribution was chosen as it is more robust toward extreme outliers than the normal distribution.

Of course, the pretraining step can extract information from the PSRK predictions only for those solutes and solvents that can in general be modeled by PSRK, i.e., for which at least one $\ln H_{ij}$ within the considered matrix can be predicted. With the present public version of PSRK,[27] this is the case for 81 of the 101 studied solutes (80.2%) and 142 of the 247 studied solvents (57.5%). Hence, only for those 81 solutes and 142 solvents, meaningful preliminary LVs can be inferred from the PSRK predictions and, as a consequence, an informative prior for the subsequent refinement step can be generated. For those solutes and solvents that cannot be modeled by PSRK, the same uninformative prior as for the training of the MCM-data was chosen in the refinement step of MCM-hybrid: a normal distribution centered around zero with standard deviation $\sigma_{\mathrm{P}} = 1$ for $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ and $\sigma_{\mathrm{P,CB}} = 10$ for b_i^u and b_j^v.

For those solutes and solvents that can be modeled by PSRK, an informative prior for the LVs in the refinement step was generated from the posterior of the pretraining step as described in the following. Since the posterior of the pretraining step of the studied LVs was approximately normally distributed in all cases, they were fitted with normal distributions yielding mean and standard deviation for each LV. The means were adopted, whereas the standard deviations of all informed solute and solvent biases were subsequently scaled with a constant factor, such that the mean of all resulting standard deviations was $\sigma_{\mathrm{P,CB}} = 5$; similarly, the standard deviations of the remaining informed LVs were scaled to yield a mean standard deviation of $\sigma_{\mathrm{P}} = 0.5$. The scaling factors, which can be seen as hyperparameters, were set to 6.44 for σ_{P} and 172.08 for $\sigma_{\mathrm{P,CB}}$, respectively, to obtain the specified mean standard deviations. This scaling procedure is necessary, since the predictions of PSRK are in general less trustworthy than the experimental data, and show some extreme outliers as exemplified in Figure S5. Without this scaling, the predictions of PSRK and the experimental data would be basically treated in the same way, resulting in an exaggerated influence of the PSRK predictions on the training of the hybrid MCM. By setting the mean standard deviation to half of the values for the uninformed prior ($\sigma_{\mathrm{P}} = 1$ or $\sigma_{\mathrm{P,CB}} = 10$, cf., above), a stronger prior was obtained for those LVs for which a priori information could be extracted from the PSRK predictions. However, these informed prior probability distributions for the LVs are still broad enough to give enough flexibility in the refinement step, if sufficient evidence is provided by the experimental training data.

The scaling of the posterior distributions from the pretraining step can in general lead to distributions that are broader than the uninformed prior. Therefore, a last processing step was introduced to ensure that the informed prior for those solutes and solvents for which a priori information could be extracted from the PSRK predictions is *always* stronger than the uninformed prior for those solutes and solvents for which this is not the case. This was achieved by multiplying the scaled posterior from the pretraining step with the respective uninformative prior distributions, resulting in the final informative prior for the refinement step of MCM-hybrid:

$$p(u_{i,k}) = \mathcal{N}\left(u_{i,k}^{*}, \sigma_{\mathrm{P}}^{*}\right), \quad \text{for } k = 1 \ldots K \tag{12}$$

$$p(v_{j,k}) = \mathcal{N}\left(v_{j,k}^{*}, \sigma_{\mathrm{P}}^{*}\right), \quad \text{for } k = 1 \ldots K \tag{13}$$

$$p\left(b_i^u\right) = \mathcal{N}\left(b_i^{u*}, \sigma_{\mathrm{P,CB}}^{*}\right) \tag{14}$$

$$p\left(b_j^v\right) = \mathcal{N}\left(b_j^{v*}, \sigma_{\mathrm{P,CB}}^{*}\right). \tag{15}$$

Again, normal prior distributions were used for all LVs, but not centered around zero (as in the pretraining step of MCM-hybrid and the training step of MCM-data) but centered around an initial guess for each LV ($\boldsymbol{u}_i^{*}, \boldsymbol{v}_j^{*}, b_i^{u*}, b_j^{v*}$) based on the posterior of the preceding pretraining step; also the standard deviations of the prior distributions (σ_{P}^{*}, $\sigma_{\mathrm{P,CB}}^{*}$) in the refinement step were set based on the posterior of the preceding pretraining step.

The final prior (informative for components that can be modeled with PSRK, uninformative for components that cannot be modeled with PSRK) was ultimately used in the refinement step of MCM-hybrid, in which the method was trained to the available experimental data for $\ln H_{ij}$ from the DDB, cf., Figure 1. In the refinement step, a normal likelihood with standard deviation $\sigma_{\mathrm{L}} = 0.2$ was chosen, which is in analogy to the (single) training step of MCM-data:

$$\begin{aligned} p\left(\ln H_{ij} | \boldsymbol{u}_i, \boldsymbol{v}_j, b_i^u, b_j^v\right) &= \mathcal{N}\left(\boldsymbol{u}_i^{\mathrm{T}} \cdot \boldsymbol{v}_j + b_i^u + b_j^v, \sigma_{\mathrm{L}}\right) \\ &= \mathcal{N}\left(u_{i,1} \cdot v_{j,1} + \cdots + u_{i,K} \cdot v_{j,K} + b_i^u + b_j^v, \sigma_{\mathrm{L}}\right). \end{aligned} \tag{16}$$

Similar to MCM-data, preliminary studies have shown that MCM-hybrid exhibits robust behavior for hyperparameters over a wide range. The procedure can be adapted as needed, but the one proposed here works well for predicting Henry's law constants.

4 | COMPUTATIONAL DETAILS

Both MCMs introduced in this work were implemented in the probabilistic programming language *Stan*[37]; details on the models including the source code to run them in Stan are given in Figures S2–S4. As inference method, which inverts the generative process of the probabilistic model and reasons about the LVs for given data ($\ln H_{ij}$ here), we resorted to Gaussian mean-field variational inference,[36] which approximates the posterior probability densities by solving an optimization problem.[35] Since exact Bayesian inference is intractable except for very simple cases and usually no closed form solution is accessible, variational inference is commonly employed for this purpose and has been successfully applied to various models up to large scales.[38] By sampling from the approximated posterior distributions, the distribution of the ln H_{ij} numbers can be calculated for any combination of solutes and solvents from Equation (1). All pre- and postprocessing steps were performed in MATLAB® R2019b.[39]

For evaluating the predictive performance of the two MCMs, a leave-one-out analysis[40] was used. Each MCM was thereby trained multiple times, and in each run, one of the 2661 available experimental data points was withheld during the training and subsequently predicted by the MCM; the prediction was then compared to the withheld experimental value. This leave-one-out analysis requires that for each solute and each solvent at least two data points for different systems are available. If this condition is satisfied, there is at least one data point for each component in the training set (after withholding the test data point), such that the MCMs have at least some information for learning the features of each component. Considering the deviations between prediction and experimental value of all available data points, the overall scores *mean absolute error* (MAE) and *mean squared error* (MSE) were calculated and compared among the MCMs and with those from PSRK. The latter comparison is, however, not trivial: both MCMs developed here allow for the prediction of ln H_{ij} for *all possible* combinations of the studied solutes and solvents and therefore for notably more binary systems within the considered matrix than PSRK. They are assessed using leave-one-out analysis, i.e., based on real predictions, since the respective data point was not used for training the MCMs. In contrast, the deviations reported for PSRK are simply those from the trained method as it is reported in the literature.[27] Unfortunately, the training set that was used for obtaining the parameters of PSRK has not been disclosed in the literature. It may be speculated that it contained a large fraction of the data points that are considered here. Hence, even though we use formally the same statistical quantities to characterize the deviations for the MCMs on the one side and PSRK on the other, they refer to different types of deviations. Of course, in contrast to the MCMs, PSRK can as a group-contribution method be used to describe additional components besides the 101 solutes and 247 solvents considered here.

In the Supporting Information, we report “final” LVs for all solvents and solutes. They were inferred by MCM-hybrid using *all* 2661 experimental data points for ln H_{ij} (without applying a leave-one-out strategy). The idea behind this is to obtain a single set of parameter values that enables a direct application of the MCM for predicting ln H_{ij}. Comparing the numbers for the LVs reported in the Supporting Information and those obtained in the leave-one-out analysis reveals, as expected, only minor differences. Consequently, the numerical values in the Supporting Information constitute a *complete* parameter set of the *final* MCM-hybrid model and allows the prediction of ln H_{ij} for any binary combination of the studied solutes and solvents at 298 K.

5 | RESULTS AND DISCUSSION

In Figure 4, the performance of the two developed matrix completion methods (MCM-data and MCM-hybrid) for the prediction of Henry's law constants ln H_{ij} in binary systems of a solute *i* and a solvent *j* at 298 K is evaluated in terms of MAE and MSE and compared to the performance of PSRK.[27] As described above, the scores of the MCMs are thereby obtained by a leave-one-out analysis and comparing the predictions with the respective experimental data from the DDB.[32] Note that in addition to the two MCMs discussed here, a third MCM was tested. It is a variant of MCM-data but without considering component biases, cf., Equation (1). The results are presented in the Supporting Information and show that using the biases yields substantially better results.

In Figure 4, only those data points that can be described with PSRK are considered. By using the latest published parameterization given by Horstmann et al.,[27] PSRK can predict ln H_{ij} for 1438 of the 2661 binary systems (54.0%) for which experimental data are available in the DDB,[32] cf., Figure 1.

The results in Figure 4A show that the MAE and MSE of PSRK are substantially larger than the respective scores of both MCMs (note the logarithmic scale). However, a closer analysis shows that the poor scores of PSRK can mainly be attributed to only a handful of data points that are extremely badly predicted by PSRK, as exemplified in Figure S5. Most of these extreme outliers correspond to the solute hydrochloric acid (HCl) dissolved in alcohols as solvents and can be attributed to poor group-interaction parameters between the HCl group and the alcohol group of PSRK. To obtain a fairer comparison, we have also omitted these extreme outliers for calculating the MAE and MSE of the methods and represent the respective scores in Figure 4B.

When omitting the PSRK outliers, the performance of MCM-data is similar to that of PSRK. The hybrid approach MCM-hybrid clearly outperforms PSRK and MCM-data in both scores irrespective of whether the PSRK outliers are taken into account or not. It is interesting to realize that MCM-hybrid, which combines information from PSRK predictions with scarce experimental data in the training, apparently does not suffer from the extreme PSRK outliers. This underpins the robustness of MCM-hybrid. The results shown in Figure 4 also demonstrate that the Bayesian approach for the hybridization works well and combines advantages of PSRK with those of the data-driven MCM, while not being impaired by the weaknesses of the individual methods.

In Figure 5, the predictions of PSRK, MCM-data, and MCM-hybrid are compared in a parity plot (panel A) and a histogram representation of the deviations from the experimental data (panel B).

The representations in Figure 5 support the findings described above. Figure 5A clearly indicates that the hybrid approach particularly improves the prediction of those data points that are rather poorly predicted with PSRK or MCM-data (or both), which is consistent with the observation of a substantially lower MSE in Figure 4. This again indicates that MCM-hybrid represents an extremely robust combination of two approaches that benefits from additional information but is not notably prone to shortcomings of the individual methods. Furthermore, Figure 5B illustrates that MCM-hybrid predicts most data points with a very high accuracy; the deviations are often in the range of $\left|\Delta \ln\left(H_{ij}/\mathrm{kPa}\right)\right| < 0.1$, corresponding to

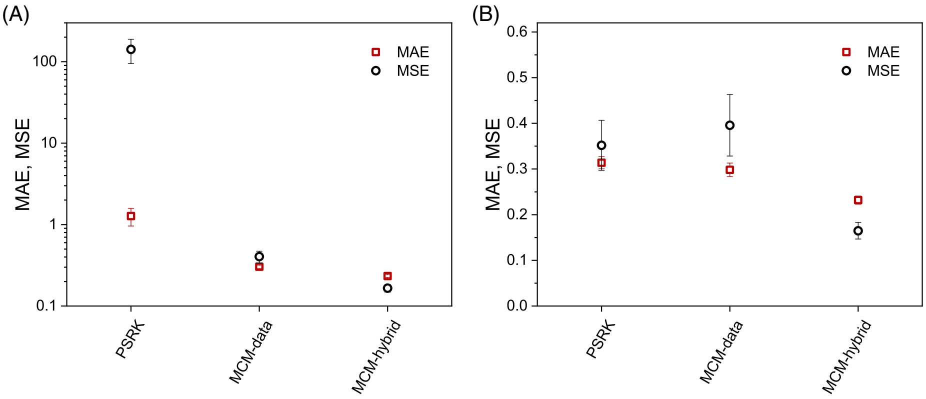

FIGURE 4 Mean absolute error (MAE) and mean squared error (MSE) of PSRK, MCM-data, and MCM-hybrid for the prediction of lnH_{ij} for binary systems at 298 K. (A) Considering the full data set (1438 data points). (B) Without considering the worst 11 outliers of PSRK, cf., Figure S5.

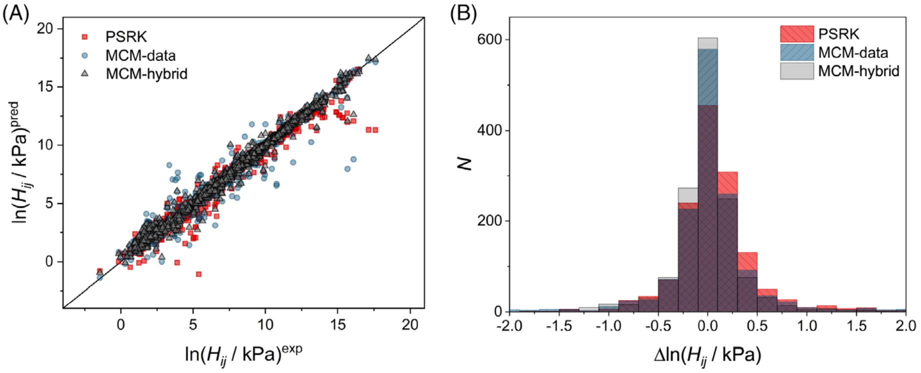

FIGURE 5 Comparison of the predictions (pred) for lnH_{ij} with PSRK, MCM-data, and MCM-hybrid without considering the worst 11 outliers of PSRK. (A) Parity plot of predictions over experimental data (exp) from the DDB. (B) Histogram of the deviations of the predictions from the experimental data. N is the number of binary systems. The shown interval in the histogram contains 97.8% (PSRK), 98.4% (MCM-data), and 99.5% (MCM-hybrid) of all considered data points.

deviations that are in the order of the experimental uncertainty in the determination of Henry's law constants. For instance, we have estimated the experimental uncertainty of ln H_{ij} by calculating the mean standard deviation for those binary systems for which multiple data points in the temperature range of 298.15 ± 1 K were available in the DDB and found a value of almost exactly 0.1.

Unlike the proposed MCMs, PSRK is, as group-contribution method, also able to model systems outside the considered matrix, which is not the case for all MCMs presented here. However, one major disadvantage of PSRK is that its application is limited to those components and systems for which the method has been parameterized. As described above, only about 54.0% of the experimental data on H_{ij} taken from the DDB in this work can be predicted with the present public version of PSRK, which is why the comparison in Figures 4 and 5 was only carried out based on those 54.0% of the data points. This restriction does not apply for the MCMs developed in this work, as they allow the prediction of H_{ij} of *all* possible binary systems of the considered solutes and solvents. This enables the evaluation of the predictive performance of the MCMs based on all 2661 available experimental data points for H_{ij} by leave-one-out analysis, which is discussed in the following.

Figure 6 shows the MAE and MSE of the predictions with MCM-data and MCM-hybrid; Figure S11 depicts the predictions of both methods in a parity plot (panel A) and a histogram representation (panel B) similar to Figure 5. In Figure S12, a parity plot that additionally includes information on the model uncertainties is given.

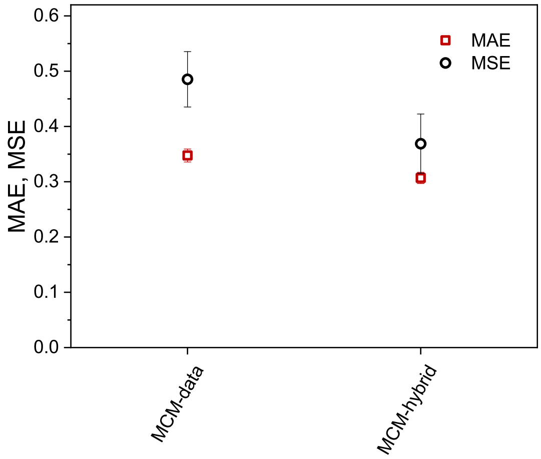

FIGURE 6 Mean absolute error (MAE) and mean squared error (MSE) of MCM-data and MCM-hybrid for the prediction of lnH_{ij} in binary systems at 298 K. For the evaluation, *all* 2661 experimental data points from the DDB were considered here.

The observations are similar to those discussed above. The scores are slightly worse when all available data are considered than when only data that can be modeled with PSRK are considered. This is not unexpected, as those components that cannot be described with PSRK are in general less studied, i.e., for these components, less data for training the MCMs are available.

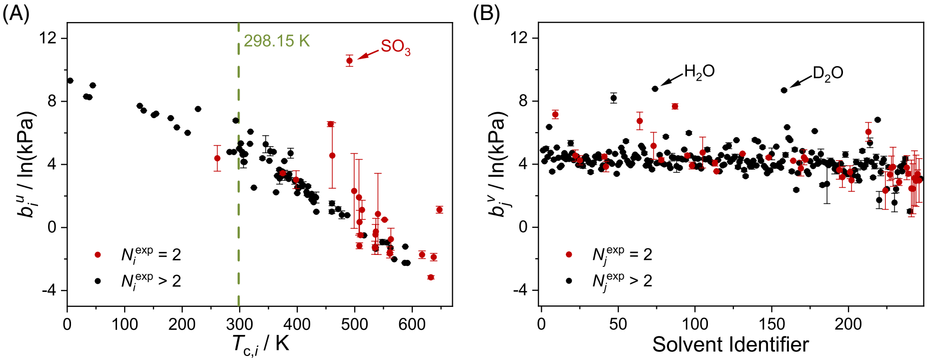

FIGURE 7 Component biases of all solutes (A, ordered according to the critical temperature) and solvents (B, ordered according to the DDB number) as inferred by MCM-hybrid. Means (symbols) and standard deviations (error bars) were calculated from the results of the leave-one-out runs assuming normal distributions for the predictions. Solutes and solvents for which only data for two different systems are available in the data set are marked red.

In the following, we briefly discuss how MCMs can not only be applied for the prediction of mixture properties (ln H_{ij} here), but also enable interesting physical insights in the mixture data. We therefore study the LVs of the solutes and solvents that were inferred during the training of MCM-hybrid from the mixture data (PSRK predictions and experimental data on ln H_{ij}) in more detail.

Figure 7 shows the component biases b_i^u and b_j^v of all solutes (panel A) and all solvents (panel B), respectively; the solutes and solvents are ordered in analogy to Figure 1, i.e., the solutes are sorted according to their critical temperature in ascending order, while the solvents are arranged by their DDB number (which is rather arbitrary). Similar figures for the remaining LVs ($\boldsymbol{u}_i$ and $\boldsymbol{v}_j$) are shown in Figures S13 and S14.

The number of data points that was considered for each solute (solvent) in Figure 7 equals the number of different binary systems in the data set that contain the respective solute N_i (solvent N_j). This number of data points is attributed to the performed leave-one-out analysis, where one experimental ln H_{ij} was withheld in each run and all LVs were trained. Thereby, only those LVs were saved that were obtained when the considered solute (solvent) was part of the one system that was withheld. From the selected data points, mean and standard deviation of b_i^u (b_j^v) were calculated and are depicted as symbols and error bars in Figure 7, respectively. While $N_i = 2$ and $N_j = 2$ often lead to high standard deviations, rather small standard deviations are observed for most solutes and solvents that appear at least three times in the data set, i.e., for $N_i \geq 3$ and $N_j \geq 3$, respectively.

In our previous work,[7,8] in which we employed MCMs for the prediction of activity coefficients at infinite dilution, no component biases were used. This is motivated by the fact that there is no such thing as a solute that exhibits *in general* small (or *in general* large) activity coefficients in *any* solvent, and, analogously, there is no solvent that *in general* leads to small (large) activity coefficients of *any* solute. By contrast, there are, for example, gases whose solubility is *in general* rather small (or large) *irrespective* of the solvent, and we take this fact into account by considering component biases for the prediction of Henry's law constants here; of course, a single gas does not exhibit the *exact same* solubility in all solvents, which we take into account by the other LVs that are considered. For the solute bias b_i^u, a clear correlation with the solute's critical temperature $T_{c,i}$ is found: b_i^u decreases with increasing $T_{c,i}$, cf., Figure 7A. This is consistent with Figure 1 and the expectation that solutes with high critical temperatures generally have a higher solubility than solutes with low critical temperature. For instance, helium and hydrogen, which have a very low critical temperature, are quite poorly soluble irrespective of the considered solvent. For the solvent bias b_j^v, no trend and only rather small variations are found (except for "extreme" molecules like water and heavy water), cf., Figure 7B, which supports the hypothesis discussed in the analysis of Figure 1 that the type of solute has a stronger influence on H_{ij} than the type of solvent. These observations do not only allow interesting physical insights, but also open the path for an estimation of the solute and solvent biases of components that are not included in the current data set. For instance, b_i^u could roughly be estimated from $T_{c,i}$ using the correlation shown in Figure 7A, whereas for b_j^v, the average value of all solvent biases depicted in Figure 7B could be used. The situation is more complicated when the other LVs are considered, cf., Figures S13 and S14. However, the examples shown in Figure 7 underline that correlations of the LVs with physical descriptors can be found, even though they may not be as simple as in these fortunate cases.

6 | CONCLUSIONS

In the present work, we have introduced a new class of prediction methods for Henry's law constants H_{ij}, namely matrix completion methods (MCMs), and have demonstrated their applicability for H_{ij} of solutes i in pure solvents j at 298 K. The idea behind this approach is that binary data can conveniently be stored in a matrix and that MCMs, which are well established in machine learning, can be applied for completing matrices even in cases where they are only sparsely occupied with experimental data, as it is the case for H_{ij} (and many other mixture properties). Two MCMs for predicting H_{ij} were implemented in the present work using a Bayesian framework and the probabilistic programming language Stan. The first MCM is purely data-driven, i.e., it is trained only to the scarce available experimental data on H_{ij}, while the second MCM follows a hybrid approach by additionally incorporating predictions from the Predictive Soave-Redlich-Kwong (PSRK) equation-of-state. The performance of both MCMs for predicting H_{ij} for 101 solutes i and 247 solvents j was evaluated by a leave-one-out analysis using experimental data from the Dortmund

Data Bank (DDB).[32] While with the purely data-driven MCM a predictive accuracy comparable to that of PSRK was found, a substantially better performance was obtained with the hybrid MCM.

The introduced MCMs have broad applicability: they are capable of predicting the H_{ij} for all 24,947 possible binary systems of the considered solutes and solvents as they are not limited by unavailable parameters; in contrast, PSRK can only predict $\ln H_{ij}$ for 31.1% of these binary systems. Of course, as group-contribution method, PSRK can in principle also be applied for predicting H_{ij} in systems containing other solutes and solvents than those studied here. Furthermore, while the refinement and extension of physics-based prediction methods like PSRK is very elaborate, the MCMs presented in this work can be adapted in a straightforward manner when additional data become available. Moreover, the presented matrix completion approach is not restricted to the prediction of Henry's law constants but can be transferred to other thermodynamic properties in a straightforward manner. The success of the MCMs is thereby based on uncovering structure in the respective mixture data, which can also be expected for many other thermodynamic properties. Also refinements and extension of the MCMs through the additional incorporation of pure component descriptors is an exciting field for future research. The same holds for considering the temperature dependence of the Henry's law constants.

AUTHOR CONTRIBUTIONS

Nicolas Hayer: Conceptualization (equal); data curation (lead); investigation (lead); writing – original draft (lead). **Fabian Jirasek:** Conceptualization (equal); project administration (lead); supervision (equal); writing – review and editing (equal). **Hans Hasse:** Conceptualization (equal); funding acquisition (lead); supervision (equal); writing – review and editing (equal).

ACKNOWLEDGMENT

We gratefully acknowledge financial support by Carl Zeiss Foundation in the frame of the project "Process Engineering 4.0" and by Bundesministerium für Wirtschaft und Energie (BMWi) in the frame of the project "KEEN".

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Nicolas Hayer https://orcid.org/0000-0002-7321-8532
Fabian Jirasek https://orcid.org/0000-0002-2502-5701

REFERENCES

1. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Found Comput Math*. 2009;9:717-772.
2. Parhizkar R, Karbasi A, Oh S, Vetterli M. Calibration using matrix completion with application to ultrasound tomography. *IEEE Trans. Signal Process*. 2013;61:4923-4933.
3. Ledent A, Alves R, Kloft M: Orthogonal inductive matrix completion. *arXiv e-prints*. 2020; arXiv:2004.01653.
4. Teflioudi C, Makari F, Gemulla R: Distributed matrix completion. 2012 IEEE 12th International Conference on Data Mining; 2012; 655–664.
5. Ramlatchan A, Yang M, Liu Q, Li M, Wang J, Li Y. A survey of matrix completion methods for recommendation systems. *Big Data Mining Anal*. 2018;1:308-323.
6. Bennett J, Lanning S. The Netflix Prize, 2007.
7. Jirasek F, Alves RAS, Damay J, et al. Machine learning in thermodynamics: prediction of activity coefficients by matrix completion. *J Phys Chem Lett*. 2020;11:981-985.
8. Jirasek F, Bamler R, Mandt S. Hybridizing physical and data-driven prediction methods for physicochemical properties. *Chem Commun*. 2020;56:12407-12410.
9. Salakhutdinov R, Mnih A: Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Cohen W (Ed). Proceedings of the 25th International Conference on Machine Learning; 2008; ACM, New York, NY.
10. Kim Y-D, Choi S. Scalable variational Bayesian matrix factorization with side information. *Artif Intell Stat*. 2014;33:493-502.
11. Damay J, Jirasek F, Kloft M, Bortz M, Hasse H. Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion. *Ind Eng Chem Res*. 2021;60:14564-14578.
12. Ahmadi M-A, Pouladi B, Javvi Y, Alfkhani S, Soleimani R. Connectionist technique estimates H_2S solubility in ionic liquids through a low parameter approach. *J Supercrit Fluid*. 2015;97:81-87.
13. English NJ, Carroll DG. Prediction of Henry's law constants by a quantitative structure property relationship and neural networks. *J Chem Inf Comput Sci*. 2001;41:1150-1161.
14. O'Loughlin DR, English NJ. Prediction of Henry's law constants via group-specific quantitative structure property relationships. *Chemosphere*. 2015;127:1-9.
15. Wang Z, Su Y, Jin S, et al. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chem*. 2020;22:3867-3876.
16. Katritzky AR, Kuanar M, Slavov S, et al. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem Rev*. 2010;110:5714-5789.
17. Goodarzi M, Ortiz EV, Coelho LDS, Duchowicz PR. Linear and nonlinear relationships mapping the Henry's law parameters of organic pesticides. *Atmos Environ*. 2010;44:3179-3186.
18. Duchowicz PR, Aranda JF, Bacelo DE, Fioressi SE. QSPR study of the Henry's law constant for heterogeneous compounds. *Chem Eng Res Des*. 2020;154:115-121.
19. Ghaslani D, Eshaghi Gorji Z, Ebrahimpoor Gorji A, Riahi S. Descriptive and predictive models for Henry's law constant of CO_2 in ionic liquids: a QSPR study. *Chem Eng Res Des*. 2017;120:15-25.
20. Li H, Wang X, Yi T, Xu Z, Liu X. Prediction of Henry's law constants for organic compounds using multilayer feedforward neural networks based on linear salvation energy relationship. *J Chem Pharm Res*. 2014;6:1557-1564.
21. Gmehling J, Kolbe B, Kleiber M, Rarey JR. *Chemical Thermodynamics for Process Simulation*. Wiley-VCH-Verl; 2012.
22. Fredenslund A, Jones RL, Prausnitz JM. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE J*. 1975; 21:1086-1099.
23. Constantinescu D, Gmehling J. Further development of modified UNIFAC (Dortmund): revision and extension 6. *J Chem Eng Data*. 2016;61:2738-2748.
24. Dahl S, Michelsen ML. High-pressure vapor-liquid equilibrium with a UNIFAC-based equation of state. *AIChE J*. 1990;36:1829-1836.
25. Patel NC, Abovsky V, Watanasiri S. Calculation of vapor–liquid equilibria for a 10-component system: comparison of EOS, EOS–GE and GE–Henry's law models. *Fluid Phase Equilib*. 2001;185:397-405.
26. Huron M-J, Vidal J. New mixing rules in simple equations of state for representing vapour-liquid equilibria of strongly non-ideal mixtures. *Fluid Phase Equilib*. 1979;3:255-271.

27. Horstmann S, Jabłoniec A, Krafczyk J, Fischer K, Gmehling J. PSRK group contribution equation of state: comprehensive revision and extension IV, including critical constants and α-function parameters for 1000 components. *Fluid Phase Equilib*. 2005;227:157-164.
28. Holderbaum T, Gmehling J. PSRK: a group contribution equation of state based on UNIFAC. *Fluid Phase Equilib*. 1991;70:251-265.
29. Soave G. Equilibrium constants from a modified Redlich-Kwong equation of state. *Chem Eng Sci*. 1972;27:1197-1203.
30. Horstmann S, Fischer K, Gmehling J. PSRK group contribution equation of state: revision and extension III. *Fluid Phase Equilib*. 2000;167: 173-186.
31. Yang Q, Zhong C. A modified PSRK model for the prediction of the vapor-liquid equilibria of asymmetric systems. *Fluid Phase Equilib*. 2001;192:103-120.
32. Dortmund Data Bank (2019).
33. Raghuwanshi SK, Pateriya RK. Collaborative filtering techniques in recommendation systems. *Data, Engineering and Applications*. Springer; 2019:11-21.
34. Aggarwal CC. *Recommender Systems: the Textbook*. 1st ed. Springer; 2016.
35. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc*. 2017;112:859-877.
36. Kucukelbir A, Tran D, Ranganath R, Gelman A, Blei DM. Automatic differentiation Variational inference. *J Mach Learn Res*. 2017;18:1-45.
37. Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76:1-32.
38. Zhang C, Butepage J, Kjellstrom H, Mandt S. Advances in Variational inference. *IEEE Trans Pattern Anal Mach Intell*. 2019;41:2008-2026.
39. MATLAB R2019b.
40. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer; 2017, corrected at 12th printing 2017.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Hayer N, Jirasek F, Hasse H. Prediction of Henry's law constants by matrix completion. *AIChE J*. 2022;68(9):e17753. doi:10.1002/aic.17753