Fluid Phase Equilibria 549 (2021) 113206

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Perspective article

Perspective: Machine Learning of Thermophysical Properties

Fabian Jirasek, Hans Hasse*

Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 23 August 2021
Accepted 27 August 2021
Available online 29 August 2021

Keywords:
Machine learning
Thermodynamic data
Physical modeling
Hybrid methods
Matrix completion methods

ABSTRACT

In this first contribution to Fluid Phase Equilibria's Perspective Series, we discuss the role of machine learning (ML) in research on thermophysical properties. Following the idea behind the new series, this is no classical review aiming at a comprehensive coverage of previous work on the field. Instead, we provide an overview of the developments and point out promising new directions in the field, linking the perspectives of chemical engineers and computer scientists. The topics we cover include the role of data in research on thermophysical properties; the long history of ML methods in this field, which, however, stemmed so far almost exclusively from supervised learning; other ML methods of interest; as well as the important subject of how to merge physical modeling with ML to create hybrid approaches, which we expect to play a central role in the future. The discussion is illustrated by examples of the application of matrix completion methods from ML for the prediction of mixture properties, which we have recently introduced.

1. Introduction

The current enthusiasm for *machine learning (ML)* has also taken hold of the research on thermophysical properties, and we are experiencing a rapid increase of the interest and activities in this field. This has been triggered by the breathtaking progress in the development of ML algorithms and their (open access) availability, and it has been amplified by a remarkable expansion of funding opportunities.

A number of pertinent questions arise on this background:

- What exactly is the enthusiasm about? Which ML techniques are relevant for the research on thermophysical properties, or could become relevant?
- What is the novelty of applying ML in research on thermophysical properties?
  Is ML a game-changer in this field?
- Which standpoint should researchers on thermophysical properties take regarding the present ML excitement? *"Keep calm and carry on!"* – or rather *"Hooray, and up she rises!"*

We want to contribute here to the ongoing discussion of these topics; albeit not aiming at giving a review of the field.

Machine learning is about machines (computers or algorithms) that learn from data how to solve problems, and thereby gradually improve themselves by experience. We are interested here in the application of ML in research on thermophysical properties, where the most prominent application is using ML for the prediction of thermophysical properties of pure components and mixtures. Given the scope of *Fluid Phase Equilibria,* we will focus on the properties of fluids.

2. The role of data in research on thermophysical properties

Data have always played a predominant role in the research on thermophysical properties. Up to quite recently, in this context, *data* meant generally data from *laboratory experiments*. However, the importance of data generated by *computer experiments*, both from quantum chemistry and from molecular simulations based on force fields, is rapidly gaining ground. There are interesting issues about using data from computer experiments besides the laboratory data, which we will, however, not address here for brevity; and in the following, we will generally not differentiate between these different types of data.

The predominant role of data in the research on thermophysical properties can be considered as a direct consequence of the fact that the number of different pure substances is extremely large, not to speak of the number of different mixtures, and that a comprehensive, practically feasible theory is still lacking, despite the considerable advancements of quantum chemistry. Nevertheless, it is true that the data-driven research on thermophysical properties has substantially contributed to establishing some of the most fundamental laws of physics, and to the general development of physics, namely in the field of thermodynamics.

* Corresponding author.
E-mail address: hans.hasse@mv.uni-kl.de (H. Hasse).

https://doi.org/10.1016/j.fluid.2021.113206
0378-3812/

How are the data used in the research on thermophysical properties? In general, for the development of (mathematical) models of the properties; and here, basically, in two ways: firstly, to adjust model parameters, and secondly, to validate the model's structure, and to adapt it, if necessary. Here, the term *model* encompasses very different things; let us first consider the extremes: there are models that contain no physical information (e.g., a polynomial fit); they obtain their usefulness only from the adjustment of their parameters to the data. On the other side, there are models that are based on physical theory (e.g., the van der Waals equation of state), which, however, usually still contain parameters that have to be adjusted to some data points. In the continuum between these extremes, the most important feature of a model, irrespective of its type, is its *predictive capability*, i.e., the ability to give predictions for situations it has not been fitted to. It is often argued that predictive capability is a consequence of the physical theory behind the model.

3. Machine learning and modeling of thermophysical properties

Let us now look at models of thermophysical properties from an ML perspective. To begin with, we will consider only models with no (or only little) physical background. The non-linear parameter fitting, which is commonly used for their development, is a regression problem and can, as such, be considered as an ML problem. The ML world is often grossly divided in the fields: *supervised learning, unsupervised learning,* and *reinforcement learning*, though there are also intermediate classes, such as *self-supervised learning* and *semi-supervised learning* [6,13,15,18,27]. Fitting the parameters of a model to data by minimizing a specified loss function belongs to the field of supervised learning. Generally, most ML techniques that have been applied so far in the research on thermophysical properties can be assigned to this field; we will therefore focus on supervised learning in the following and only briefly discuss the prospective role of other ML techniques for the research on thermophysical properties at the end of this communication.

3.1. The workhorse – supervised learning

Supervised learning is about inferring a relation between input and output values based on a set of *labeled* training data (where labeled refers to the fact that input-output pairs are known and given) and using this relation for predicting the output of new data points given their input values. This is exactly the problem we face in developing models of thermophysical properties. Supervised learning includes not only fitting the parameters of the model, but also finding a suitable structure of the model, i.e., the model selection. All this requires solving optimization problems. Machine learning offers methods and a very large and rapidly expanding toolbox for supervised learning, starting with linear regression, support vector regression, and Gaussian process regression and going way to deep neural networks [18], to name only a few classes of algorithms as example. And ML deals with fundamental aspects of the learning process itself. Practitioners are aware of the risks of *overfitting*, i.e., using too many parameters such that the training data are well described but leading to unreliable extrapolations (and eventually even unreliable interpolations), as well as the problems resulting from underfitting, i.e., not having enough parameters to obtain a good description of the data. In ML, this is called the *bias-variance dilemma* [5]; and there is extensive literature in which it is systematically discussed, including not only the influence of the amount and quality of training data but also strategies to detect and handle it in practice, e.g., by cross-validation or regularization [6,18].

Machine learning thereby builds on well-known mathematical concepts from optimization and statistics, complemented by novel developments of generic approaches such as the kernel trick [6], and produces new methods and tools at a breathtaking pace. It is highly remarkable and laudable that in the ML community, together with the publication of new methods, practically always also the corresponding tools are supplied open source. This, together with the progress in computer technology, is considerably facilitating the use of ML methods by practitioners from other fields, such as our community.

We now come back to the questions from the start: yes, many ML techniques from supervised learning are directly relevant for our field of research; and we can directly use tools developed by the ML community for modeling thermophysical data. However, it must be stated clearly that researches from our field have been using sophisticated methods from supervised learning for decades, e.g., [1]; and they have adapted them to their needs in a long-standing effort, leading to abundant experience on strengths and weaknesses of the approaches. Hence, using more recent ML techniques of supervised learning for modeling thermophysical data is not inherently interesting; it only becomes interesting if the new methods can beat the well-established methods from our field; or, even better, if, with the new methods, we can do things that we could not do before, or have not thought about doing. However, there is no point in boasting a successful application of an ML technique without comparing the results to those from a (conventional) benchmark, if available. To put in bluntly, the fact that, e.g., an a*rtificial neural network (ANN)* is used for modeling thermophysical data does not make a paper interesting by itself.

Besides the *regression problems*, which we have discussed so far and which are widespread in our research field, also *classification problems*, in which the output is *categorical*, can be tackled by supervised learning. Obviously, there are also many classification problems in the field of thermophysical properties, such as classifying mixtures according to their phase behavior or defining limits of the applicability of a certain model. One of the, admittedly many, hot topics in the present ML research is *anomaly detection* [2,16], which is another classification problem that is also relevant in our field, e.g., for experimentalists. In the research on thermophysical properties, classification problems have been tackled so far typically by proposing some ad hoc scheme, or by applying simple formal criteria. Machine learning toolboxes offer a variety of sophisticated methods for solving such problems, e.g., *k-nearest neighbours algorithms, support vector machines*, or d*ecision trees* [14,17].

3.2. Merging the worlds – hybrid models

So far, we have only discussed essentially data-driven models that do not contain physical theory, which is basically the default way of using ML methods, in particular in the areas where they are most commonly used. This way of applying ML methods is therefore fairly straightforward. Things become much trickier – but also more interesting – when we switch from these data-driven models to *hybrid models*, which incorporate *physical knowledge*. The limits of what is considered as incorporating physical knowledge are blurred; here, we refer to directly incorporating physical knowledge, and in particular physical theory and laws, in the model, and not just using the physical information encoded in thermophysical data to which the model is trained itself. There are many different terms for this type of models, besides *hybrid* also, e.g., *physics-guided, physics-infused*, and *physics-informed* are often used [19,26].

Although most classical physical models can be considered as belonging to this class, at least as long as some parameters are fitted to data, the incorporation of (physical) domain knowledge into modern ML algorithms is still in its infancy [23,26]. Obvi-

ously, finding suitable ways for doing this depends on the nature of the domain knowledge. In our field or research, this domain knowledge is extraordinarily diverse, starting from basic information on the substances, such as their molecular mass or chemical structure (i.e., simple pure component descriptors), to general rules and physical laws, such as the one that all substances become ideal gases if the density approaches zero. It is essential that future ML methods in our field yield results that intrinsically comply with these rules and laws. One could argue that they are inherently encoded in the data and an ML algorithm should simply learn them during the training to these data, but this is far overstretched. Thermodynamic theory is the result of efforts of generations, its development is a masterpiece of the human mind. *Re*-inventing thermodynamic theory based on data alone is not only unnecessary, it is way beyond anything that is presently imaginable for machines to do. Therefore: if we want the physical laws to be obeyed by our ML models, we should better build them in hard-wired. More practically, we can argue that if we already know these laws (or at least some of them), why shouldn't we use them *explicitly* for the development of ML models, instead of solely relying on physical training data, which is, moreover, often scarce and subject to uncertainties? No matter which argument we follow, creating hybrid approaches will be one of our key tasks in the research on thermophysical properties.

An interesting point here is that any physical model that is used as a benchmark for an ML method can, at the same time, be considered as domain knowledge that can be incorporated in the ML method. Suitably combining physical knowledge with ML methods requires individual analysis of problems, so there are plenty of opportunities for future research on hybrid models. However, also this is far from being a totally unexplored issue in our field: just take the abundant work on *Quantitative Structure Property Relationships (QSPR)* [20,22] as an example, in which mathematical correlations between physical descriptors (typically: pure component descriptors) and thermophysical properties have been established.

3.3. Overlooked beauties? – unsupervised learning and reinforcement learning

Let us now briefly address possible applications of unsupervised learning and reinforcement learning in the research on thermophysical properties, ML techniques that were, up to now, less frequently used in this field. Unsupervised learning is about finding structures in *unlabeled* data [6] (where unlabeled refers to the fact that no input-output pairs are given). Consider, e.g., a pure component data bank that contains entries for some pure component properties. An unsupervised learning algorithm could be used for identifying structures in these data, e.g., by clustering them into groups of components that are *similar*, without requiring a human to explicitly define similarity. An interesting question could be how these clusters determined by ML map to the chemical nature of the components. This could not only generate new physical insights, it could also be exploited for the development of improved supervised learning methods for thermophysical properties. Another obvious application of unsupervised learning in the research on thermophysical properties is *outlier detection* or *dimensionality reduction* of the data, e.g., by p*rinciple component analysis* [6].

In reinforcement learning, an agent (e.g., a robot or the control system of a chemical plant) acts autonomously in some environment that it can modify by certain actions. From time to time, the robot gets a feedback about the changes its actions have produced, which is usually labeled as *reward* and specified by a reward function. By suitably defining the environment, the set of actions the robot can take, and the reward function, the robot learns automatically, by trial and error, how to best obtain the desired results. The reward for certain actions is thereby typically *delayed* since an assessment of success can usually only be done after follow-up actions; reinforcement learning therefore requires backtracking chains of decisions [25]. Reinforcement learning could, e.g., be used for finding suitable building blocks for complex equations for describing thermophysical properties, such as multi-parametric equations of state.

4. Applications of machine learning of thermophysical properties

The number of research papers on thermophysical properties that explicitly claim to use ML methods has tremendously increased in the past decade. Many of these papers concern using ANNs for describing thermophysical data, often for properties and substances for which physical modeling is difficult or infeasible. The fundamental challenge in using supervised learning techniques, including ANNs, is that sufficient labeled training data has to be available. Unfortunately, the data on thermophysical properties are notoriously spotty; especially when only special properties or special classes of substances and, in general, when mixtures are considered. Many applications of ANNs for modeling thermophysical data have suffered from this. But there are also examples for the successful use of ML techniques for learning from such sparse data sets. We have recently achieved this by using a technique from unsupervised learning, m*atrix completion methods (MCMs)*, which are known from recommender systems [18], for the prediction of thermophysical properties of binary mixtures [11,12]. The idea behind using MCMs for this purpose is that, for fixed conditions, the data on a given property of binary mixtures can be arranged in the form of a matrix, with rows and columns corresponding to the components. These matrices are usually only sparsely occupied by measured mixture data. An example is shown in Fig. 1, which represents the available experimental data on activity coefficients γ_{ij}^{∞} of pure solutes i at infinite dilution in pure solvents j at 298 K as reported in the Dortmund Data Bank 2019 [4].

MCMs can fill the gaps in these matrices, i.e., predict the properties of unmeasured mixtures, and can thereby achieve even higher accuracies than the current physical reference methods [8,11]; of particular interest are hybrid approaches that combine MCMs with the physical methods [12]. Two examples for results of the prediction of experimental data with MCMs are shown in Fig. 2: activity coefficients at infinite dilution γ_{ij}^{∞} (left) and Henry's law constants H_{ij} (right). Results from three methods are com-

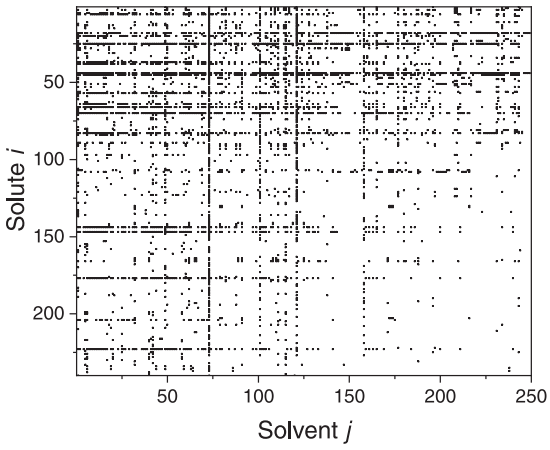

Fig. 1. Matrix schematically representing the available experimental data on activity coefficients of pure solutes i at infinite dilution in pure solvents j at 298 K as reported in the Dortmund Data Bank 2019 [4]. Black symbols denote mixtures for which data are available, white symbols denote unstudied mixtures.

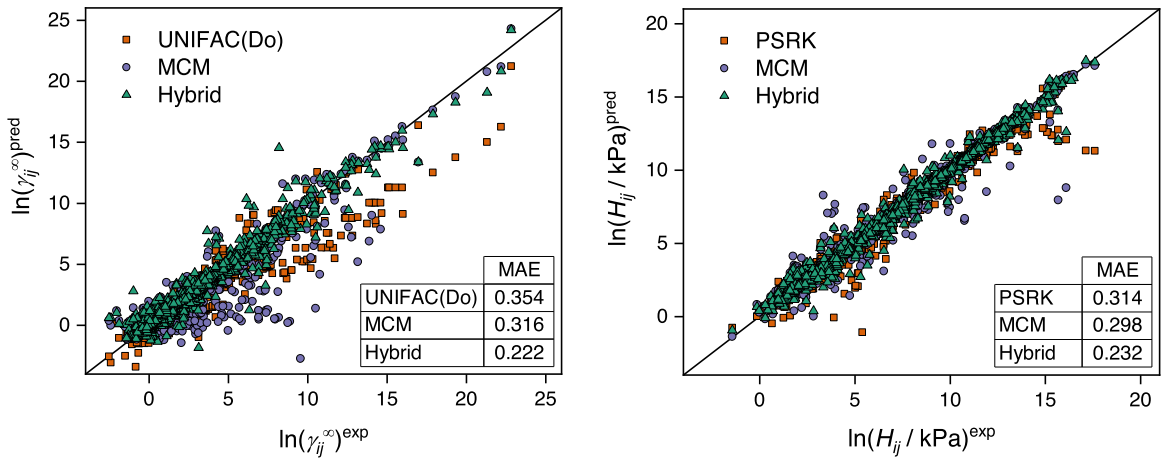

Fig. 2. Predictions (pred) of activity coefficients at infinite dilution γ_{ij}^{∞} [11,12] (left) and Henry's law constants H_{ij} [8] (right) of pure solutes i in pure solvents j at 298 K with data-driven matrix completion methods (MCM), physical reference models, namely modified UNIFAC (Dortmund) (UNIFAC(Do)) [3,24] or the Predictive Soave-Redlich-Kwong (PSRK) equation of state [9,10], and hybrid physics-based/data-driven approaches over the experimental data (exp) from the Dortmund Data Bank [4]. Inset: mean absolute error (MAE) of the predictions of the methods.

pared to experimental data: 1) physical reference methods (modified UNIFAC (Dortmund) for γ_{ij}^{∞} and the Predictive Soave-Redlich-Kwong (PSRK) equation of state for H_{ij}); 2) purely data-driven MCMs; and 3) hybrid MCMs, which combine the data-driven MCM with the respective physical method in a Bayesian framework.

5. Conclusions and outlook

Machine learning builds on well-established mathematical concepts, many of which have been used by researches in the field of thermophysical properties for many decades, without, however, calling them ML. In the research on thermophysical properties, as in the entire field of physical sciences, the fitting of model parameters to data has traditionally been considered more as an unavoidable nuisance than as a central part of the model development [7]; in ML, the adjustment of model parameters to data gets the appreciation it deserves.

As ML techniques have played an important role in our research field for decades, it is not astonishing that we can profit from the recent disruptive developments in ML research. Machine learning has much more to offer than just ANNs. Considering the modeling process from the ML standpoint gives us new perspectives. Modern ML techniques are highly relevant for our field, in particular those from supervised learning, but also unsupervised learning, reinforcement learning, and other ML techniques open new routes for the research in our field, some of which lead into unexplored territory. These routes are fairly easily accessible, as the ML community provides new ideas usually always with the tools and data needed to test them. If we were to learn nothing from the ML community but to provide models always together with tools to use them, and the raw data to which they were trained, already much would be gained.

The research on thermophysical properties differs in many ways from the fields in which ML is already routinely applied, such as picture-, text-, and speech-recognition or customer data analysis. Even the largest available data sets on thermophysical properties are small compared to those on which ML excels; and our data sets are practically always sparse and subject to considerable uncertainties. Luckily, we can outweigh this drawback by using the extraordinary knowledge on thermophysical properties that has been acquired over generations. Connecting that knowledge to ML techniques to form hybrid models is the task of the hour. The physical knowledge is the key to dealing with the problems resulting from the small size, sparsity, uncertainty, and heterogeneity of the data sets in our field. Furthermore, strategies from the *design of experiments* or *active learning* [21] should be used for devising efficient ways for extending the available data sets on thermophysical properties. We consider in particular *Bayesian ML* [18] as highly promising for research on thermophysical properties, as it intrinsically takes the uncertainties of data, parameters, and predictions into account, enables a straightforward incorporation of prior knowledge (e.g., physical laws), and can provide robust models, even for sparse data sets. All this is crucial for successfully deploying ML in our field.

Machine learning methods can not only be used in the modeling process, they are also valuable for data analysis, and they can thereby help to create physical insights and an improved understanding, both of the training data and the modeling results. This, together with the incorporation of physical knowledge, will significantly reduce the black-box character of ML tools and *create trust* and *acceptance* among users.

Machine learning of thermophysical properties is not new, but it has the potential to become a game-changer in the research on thermophysical properties – if we succeed in merging it with the abundant physical knowledge from our field. Hence, if we had to choose between: *"Keep calm and carry on!"* and *"Hooray, and up she rises!"* we would go for the *"Hooray"*, but not blindly and over-optimistically; keeping in mind where we come from, and building on the knowledge from the previous generations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We gratefully acknowledge funding from the Carl Zeiss Foundation.

References

[1] T.F. Anderson, D.S. Abrams, E.A. Grens II, Evaluation of Parameters for Nonlinear Thermodynamic Models, AIChE J. 24 (1978) 20–29.
[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A Survey, ACM Comput. Surv. (CSUR) 41 (2009) 1–58.
[3] D. Constantinescu, J. Gmehling, Further Development of Modified UNIFAC (Dort-mund): revision and Extension 6, J. Chem. Eng. Data 61 (2016) 2738–2748.

[4] Dortmund Data Bank 2019, www.ddbst.com.
[5] S. Geman, E. Bienenstock, R. Doursat, Neural Networks and the Bias/Variance Dilemma, Neural Comput. 4 (1992) 1–58.
[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2. Ed., Springer, New York, NY, 2009.
[7] H. Hasse, J. Lenhard, Boon and Bane: On the Role of Adjustable Parameters in Simulation Models, in Mathematics as a Tool, Springer, Cham, 2017.
[8] N. Hayer, F. Jirasek, H. Hasse: Prediction of Henry's Law Constants by Matrix Completion (2021) submitted.
[9] T. Holderbaum, J. Gmehling, PSRK: A Group Contribution Equation of State Based on UNIFAC, Fluid Phase Equilib. 70 (1991) 251–265.
[10] S. Horstmann, A. Jabloniec, J. Krafczyk, K. Fischer, J. Gmehling, PSRK Group Con-tribution Equation of State: Comprehensive Revision and Extension IV, Including Critical Constants and α-Function Parameters for 1000 Components, Fluid Phase Equilib. 227 (2005) 157–164.
[11] 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.
[12] F. Jirasek, R. Bamler, S. Mandt, Hybridizing Physical and Data-Driven Prediction Methods for Physicochemical Properties, Chem. Commun. 56 (2020) 12407–12410.
[13] M.I. Jordan, T.M. Mitchell, Machine Learning: Trends, Perspectives, and Prospects, Science 349 (2015) 255–260.
[14] A.I. Kadhim, Survey on Supervised Machine Learning Techniques for Automatic Text Classification, Artif. Intell. Rev. 52 (2019) 273–292.
[15] A. Kolesnikov, X. Zhai, L. Beyer, Revisiting Self-Supervised Visual Representation Learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1920–1929.
[16] D. Kwon, H. Kim, J. Kim, S.C. Suh, I. Kim, K.J. Kim, A Survey of Deep Learning-based Network Anomaly Detection, Cluster Comput. 22 (2019) 949–961.
[17] A.E. Maxwell, T.A. Warner, F. Fang, Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review, Int. J. Remote Sens. 39 (2018) 2784–2817.
[18] K. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, Cambridge, MA, 2012.
[19] R. Rau, C.K. Sahu, Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques with Cyber-Physical System (CPS) Focus, IEEE Access 8 (2020) 71050–71073 Access 8.
[20] K. Roy, S. Kar, R.N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts, Springer, 2015.
[21] B. Settles, Active Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 6 (2012) 1–114.
[22] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Comb. Sci. 22 (2003) 69–77.
[23] V. Venkatasubramanian, The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally? AIChE J. 65 (2018) 466–478.
[24] U. Weidlich, J. Gmehling, A Modified UNIFAC Model. 1. Prediction of VLE, h^E, and γ^∞, Ind. Eng. Chem. Res. 26 (1987) 1372–1381.
[25] M. Wiering, M. van Otterlo, Reinforcement Learning: State-of-the-art, Springer, 2012.
[26] J. Willard, X. Jia, S. Xu, M. Steinbach, V. Kumar: Integrating Physics-based Modeling with Machine Learning: A Survey, (2020). https://arxiv.org/abs/2003.04919.
[27] X. Zhu, A.B. Goldberg, Introduction to Semi-Supervised Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 3 (2009) 1–130.