Chemical Engineering Journal 492 (2024) 151999

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

ML-SAFT: A machine learning framework for PCP-SAFT parameter prediction

Kobi C. Felton [a,b], Lukas Raßpe-Lange [c], Jan G. Rittig [b], Kai Leonhard [c], Alexander Mitsos [e,b,d], Julian Meyer-Kirschner [g], Carsten Knösche [g], Alexei A. Lapkin [a,f,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK
[b] Process Systems Engineering (AVT.SVT), RWTH Aachen University, 52074 Aachen, Germany
[c] Institute of Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany
[d] Institute for Energy and Climate Research IEK-10: Energy Systems Engineering, Forschungszentrum Jülich GmbH, Jülich 52425, Germany
[e] JARA-ENERGY, Aachen 52056, Germany
[f] Innovation Centre in Digital Molecular Technology, Yusuf Hamied Department of Chemistry, University of Cambridge, UK
[g] BASF SE, 67056 Ludwigshafen am Rhein, Germany

ARTICLE INFO

Keywords:
Deep learning
PC-SAFT
Thermodynamics
Property predictions

ABSTRACT

The Perturbed Chain Polar Statistical Associating Fluid Theory (PCP-SAFT) equation of state (EoS) is widely used to predict fluid-phase thermodynamics, but parameterization of PCP-SAFT for individual molecules is often challenging. We propose a machine learning framework called ML-SAFT that can turn experimental data in predictive models of PCP-SAFT parameters. We demonstrate methods for automated large scale regression of PCP-SAFT parameters and thus create a large PCP-SAFT parameter dataset in the literature. We then evaluate several machine learning architectures for predicting PCP-SAFT parameters. We find that our best model provides accurate predictions for a wider range of molecules than existing predictive methods with 40 % average absolute deviation (% AAD) in vapor pressure predictions and 8 % AAD in density predictions.

1. Introduction

Fluid-phase thermodynamic predictions are required for a range of fine and bulk chemical applications, yet experimental parameterization of thermodynamic models to predict fluid-phase thermodynamics is often time and labor-intensive. This motivates the long-standing research interest in predicting parameters of thermodynamic models directly from molecular structures. In addition to established approaches such as group contribution [1–9] and quantum mechanical (QM) simulations [10–15], recent work has shown that machine learning (ML) methods can be used for predictive thermodynamics. A wide variety of ML methods have been used for predicting activity coefficients and solvation energies from molecular structures including matrix completion [16], graph neural networks [17–23], transformers [24] and a variety of other architectures [25]. However, the limitation of these works is their lack of thermodynamic consistency that comes with rigorously derived equations of state [26] or their inability to predict multiple thermodynamic properties.

To enable general and thermodynamically consistent predictions, one approach is to predict thermodynamic model parameters [27–32]. This enables straightforward use of thermodynamic models in existing process simulation software packages, which contrasts with approaches that require replacing the full thermodynamic model with a neural network [33–35]. Given the predicted parameters, the thermodynamic model can in turn be used to predict thermodynamic properties. For instance, Winter et al. [31] developed a model for predicting the parameters of the NRTL activity coefficient model for a wide range of binary mixtures.

In this work, we extend this approach of predicting parameters to an Equation of State (EoS), namely Perturbed Chain Polar Statistical Associating Fluid Theory (PCP-SAFT) [36], an established extension of the original PC-SAFT EoS to include polar molecules [37]. With the PCP-SAFT EoS, it is possible to express the residual Helmholtz free Energy as a function of the PCP-SAFT parameters. All properties computed from the PCP-SAFT EoS are derived from the Helmholtz free energy and are therefore inherently thermodynamically consistent. The advantages of PCP-SAFT include its ability to predict mixture properties using parameters regressed on pure component data (though we only explore pure component predictions in this work) and its accurate representation of polar compound properties [38].

We introduce ML-SAFT, a framework for creating machine learning models that predict PCP-SAFT parameters, shown conceptually in

* Corresponding author at: Innovation Centre in Digital Molecular Technology, Yusuf Hamied Department of Chemistry, University of Cambridge, UK.
E-mail address: aal35@cam.ac.uk (A.A. Lapkin).

https://doi.org/10.1016/j.cej.2024.151999
Received 7 May 2023; Received in revised form 23 April 2024; Accepted 6 May 2024
Available online 14 May 2024

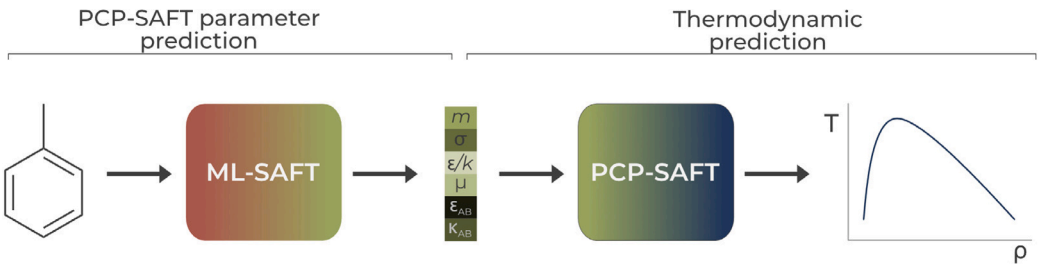

Fig. 1. ML-SAFT is a framework for predicting PCP-SAFT parameters directly from molecular structures. PCP-SAFT parameters predicted by ML-SAFT can be used in any PCP-SAFT implementation. Shown schematically is a density prediction.

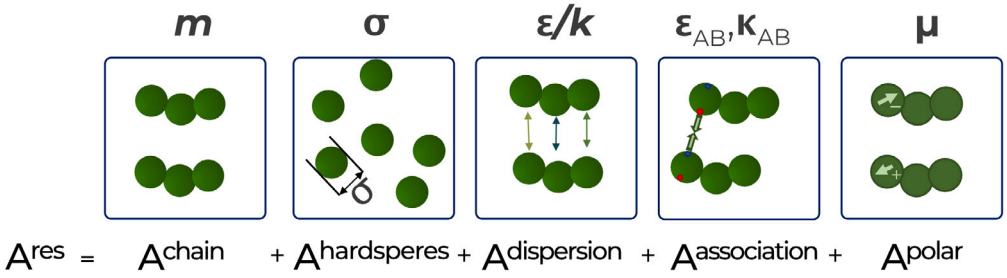

Fig. 2. Schematic illustrating the physical significance of the five PCP-SAFT parameters.

Fig. 1. The PCP-SAFT parameters are physically interpretable, but they must be regressed or predicted for each molecule. Therefore, we developed a database (871 molecules) of regressed PCP-SAFT parameters using experimental data and a combination of deep learning and heuristics. We then carried out an evaluation of several machine learning architectures for predicting these regressed PCP-SAFT parameters.

We note that Habicht et al. [39] recently developed a feed forward neural network model to predict PC-SAFT parameters from molecular fingerprints. The novelty of our framework is that we consider the complete pipeline including parameter fitting, training different ML models, and prediction of thermodynamic quantities. This enables us to generate a larger database of regressed PCP-SAFT parameters compared to the one used by [39]. Furthermore, we also consider the polar and associating terms, enabling prediction of PCP-SAFT parameters for a wider range of compounds. We also note that, concurrently to the publication of our work, Winter et al. published an ML model that embeds the PC-SAFT EoS into an ML model, specifically a Transformer operating on SMILES strings called SPT, thereby circumventing the need for parameter regression [40]. Our work offers an alternative perspective by comparing several model architectures and training approaches. Also concurrent to our publication, Esper et al. published a large database of PCP-SAFT parameters [41]. It would be interesting for future work to compare and integrate the approach by Winter et al. with other ML models that are part of ML-SAFT such as RFs and GNNs, and we would also in the future consider the larger database by Esper et al. [41].

2. Methods

2.1. The PCP-SAFT equation of state

The Perturbed-chain Statistical Associating Fluid theory (PC-SAFT) was first introduced to give an expression for the residual Helmholtz free energy [36]. The expression is based on a contribution for hard sphere interactions and a contribution for chain interactions. These two contributions describe repulsion and shape via two parameters, namely m, the number of segments in a chain of a component, and σ, the hard sphere diameter. In addition, a third term describes the strength of the dispersion and is determined by the dispersion parameter ϵ/k. These parameters are depicted in Fig. 2. For a more in depth description of the physical meaning of the PCP-SAFT parameters, we refer readers to the dissertation of Sebastian Kaminski [42].

Two terms were added to account for polar and associating effects [37,43]; this updated EoS is referred to as PCP-SAFT, which we use in this publication. The term accounting for polar interactions includes the dipole moment μ as an additional parameter. Although other multipole moments can also be used in addition, for this study, we only use the dipole moment for ML-SAFT. Finally, for the association, two more parameters are introduced: κ_{AB} is used to determine the association volume and ϵ_{AB} is used to determine the association strength. Therefore, the goal of ML-SAFT is to predict the six named parameters for a molecule necessary in order to apply the PCP-SAFT EoS. We use the implementation of PCP-SAFT in FeO_s [44].

2.2. Baseline predictive PCP-SAFT methods

There are several methods in the literature for predicting PCP-SAFT parameters. As comparisons to ML-SAFT, we evaluated two state-of-the-art methods that use QM and a group contribution method, respectively.

As a QM method, we applied the Segment-Based Equation of State Parameter Prediction (SEPP) with the 6+2+2 parameterization [14]. SEPP obtains m, σ, and ϵ/k from a multilinear model that uses DFT-calculated features as input, while the dipole moment μ is obtained directly from QM calculations. We conducted new DFT calculations for this publication. An analysis of the surface charge density from COSMO [45] was utilized to calculate the associating parameters ϵ_{AB} and κ_{AB}. In contrast to the original publication [14], we used the strongest associating site, although SEPP can take into account all binary associating interactions. This simplification was made to ensure that the predicted parameters could be used with most PCP-SAFT implementations. Since the multilinear model in SEPP was only fit to alkanes and polar compounds with oxygen and nitrogen, it is not valid for halogens, which are abundant in our dataset.

We used the homosegmented group contribution method by Sauer et al. [46] as implemented in FeO_s [44]. We used the fitted group parameters from [46] for all predictions. Some molecules in our dataset were not already segmented into groups by [46], so we used an algorithm from the python package thermo [47] to identify the groups that should be used for prediction. A modified version of the SMARTS

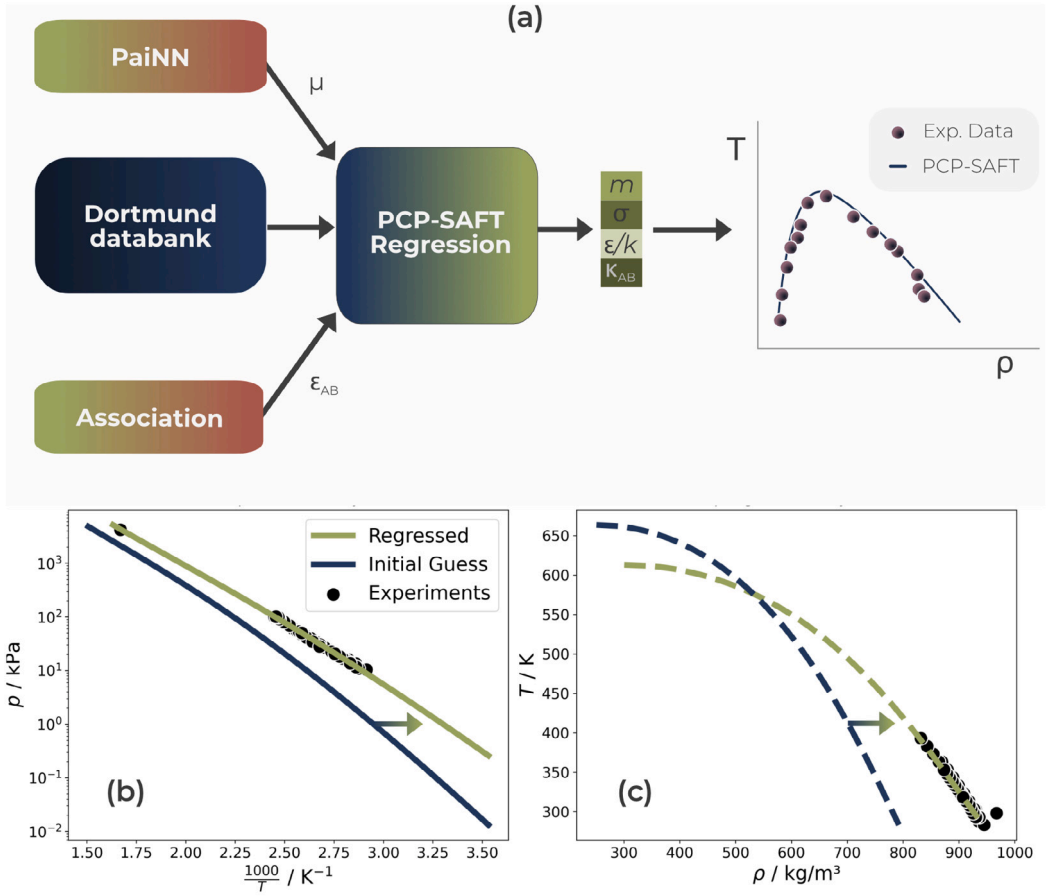

Fig. 3. Building a dataset for ML-SAFT: (a) A workflow was developed to automatically regress PCP-SAFT parameters to pure component experimental data. A machine learning model (PaiNN) trained on a combination of DFT and experimental data was used to predict the dipole moments of the experimental dataset, and the other parameters were initialized using standard values. (b–c) Example regression of PCP-SAFT to vapor pressure and density data for 2-ethoxyethanol using the Levenberg–Marquardt algorithm. The dashed line in the density plot represents liquid density.

strings from Ruggeri and Takahama (see Supplementary Data 2) were used in the fragmentation algorithm [48].

2.3. Building a dataset for ML-SAFT

2.3.1. Data extraction from the dortmund data bank

Experimental data were extracted from the Dortmund Data Bank (DDB, 2022 version) [49], which contains data for over 40k unique molecules. The software package Pura [50] was used to resolve the name or CAS numbers available in the Dortmund database into a cheminformatics friendly identifier, namely SMILES. Pura called on PubChem [51], the Chemical Identifier Resolver [52], OPSIN [53], and the Chemical Abstracts Service [54] to resolve a name or CAS number, and we required that at least two services agreed on the resolved SMILES. Pura resolved 68% (27.2k/40.3k) of names or CAS numbers to SMILES. Many of the molecules that were not resolved did not have CAS numbers or had obscure names.

The experimental data were subsequently filtered to obtain only data that were reasonable for PCP-SAFT regression. Ionic molecules were removed from the dataset as well as any molecules with temperatures outside the range 200–1000 K and pressures outside the range 10–10000 kPa. Densities greater than 2000 kg/m^3 were also excluded. Ideally, we would use the critical points to filter data, but unfortunately such data were not available in our dataset, so we simply remove the highest pressure and temperature data as a heuristic. Finally, only molecules with at least four density data points and five vapor pressure data points were considered. After all filtering steps, the experimental data for 988 unique molecules were available for regression of PCP-SAFT parameters. This significant decrease in the size of the data set from 27k to 1k by the filtering step has been noted in other attempts to build models on data available in literature databases [55,56].

2.3.2. PCP-SAFT parameters regression

We used the well-established Levenberg–Marquardt (LM) least squares algorithm and experimental vapor pressure and density data, as shown in Fig. 3. The same initial guess shown in Table 1 was applied for all molecules, which was based on the analysis of a large set of PCP-SAFT parameters calculated by QM simulation (see Section 2.2) [14]. We considered relaxing these constraints but found that wide ranges resulted in worse regression results. The following equation was applied to calculate the sum of squared errors $\mathcal{L}_i$ for molecule i:

$$\mathcal{L}_i = \sum_j \left(\frac{p_i^{sat,\mathrm{SAFT}}(T_j) - p_i^{sat,\mathrm{EXP}}(T_j)}{p_i^{sat,\mathrm{EXP}}(T_j)} \right)^2 + \sum_j \left(\frac{\rho_i^{l,\mathrm{SAFT}}(T_j, P_j) - \rho_i^{l,\mathrm{EXP}}(T_j, P_j)}{\rho_i^{l,\mathrm{EXP}}(T_j, P_j)} \right)^2 \quad (1)$$

where $p_i^{sat}(T_j)$ and $\rho_i^{l}(T_j, P_j)$ are the saturation vapor pressure and the liquid density for molecule i, respectively, at temperature T_j and P_j. The superscripts SAFT and EXP represent PCP-SAFT predictions and experimental data, respectively.

Only m, σ, ϵ/k were regressed for all molecules, and ϵ_{AB} and κ_{AB} were additionally regressed for associating molecules. Molecules were considered to be associating if they had at least one hydrogen-bond acceptor and donor site via RDKit [57]. The associating parameters

Table 1
Parameters fitted in PCP-SAFT regression to experimental data.

Parameter name	Bounds	Initial value
m	$1.0 \leq m \leq 10.0$	3.26
σ	$2.5 \leq \sigma \leq 5.0$	3.69
ϵ/k	$100.0 \leq \epsilon/k \leq 1000.0$	284
ϵ_{AB}	$0.0 \leq \epsilon_{AB} \leq 4000.0$	2400

Table 2
Data sets used to train PaiNN architecture for predicting dipole moments. μ_{source} is the method used to generate dipole moments; DFT is density functional theory, and Exp. is experimental.

Dataset	μ_{source}	Conformer type	Size	Ref.
QM9	DFT	DFT	134 k	[61]
CRC	Exp.	RDKit [60]	482	[62]
SEPP	DFT	DFT	1106	See Section 2.2

ϵ_{AB} and κ_{AB} were set to zero for these non-associating molecules. We also found that associating parameter κ_{AB} could be set to 0.01 and not regressed for all associating molecules while maintaining low regression error (see Table S1). Reducing the number of parameters to predict simplified the downstream machine learning task.

We decided to predict the dipole moment μ because previous work has shown that adjusting the dipole moment causes regression to fail due to high correlation with ϵ/k [38,58]. Specifically, we trained a deep learning model to predict dipole moments using a combination of DFT calculated and experimentally determined dipole moments, as shown in Table 2. Once trained, the model made dipole moment predictions for hundreds of molecules in seconds. For the model architecture, we chose a tensorial equivariant message passing neural network called PaiNN developed by Schütt et al. since it has been shown to give accurate predictions of dipole moment [59]. Briefly, PaiNN takes as input a relaxed conformer of a molecule and uses a series of message passing steps on both a vector and rank three tensorial representation to produce a representation of each atom. For training, we used the conformer generation methods shown in Table 2, and for inference, we used the RDKit ETKDGv3 algorithm to generate conformers [60]. Subsequently, the dipole moment was calculated using the final vector and tensorial representations of the network:

$$\vec{\mu} = \left[\sum_{i=1}^{N} \vec{\mu}_{atom}(\vec{\mathbf{v}}_i) + q_{atom}(\mathbf{s}_i)\vec{r}_i\right] \tag{2}$$

where $\mathbf{s}_i$ is the vector representation and $\mathbf{v}_i$ is the tensorial representation, $\vec{r}_i$ are the positions of the atoms and μ_{atom} and q_{atom} are both feedforward networks. Training for 63 epochs resulted in a validation mean absolute error of 0.005 for held-out dipole moment predictions.

2.4. ML-SAFT machine learning models

For prediction of the regressed PCP-SAFT parameters from molecular structures, we tested several machine learning architectures that have previously been successfully applied to molecular property prediction tasks. We included a random forest (RF) [63] and a standard feed-forward network (FFN) that use ECFP4 fingerprints with 2048 bits as input features [64]. RFs are known to have strong performance for molecular property prediction in drug discovery but are less common in process systems engineering [65,66]. Feed-forward networks were used successfully by Habicht et al. in previous work on predicting PCP-SAFT parameters [39]. Furthermore, we developed a standard message passing neural network (MPNN) [67] that has previously been used to predict several thermodynamic parameters including fuel properties [68] and activity coefficients [21,23]. MPNNs use a molecular graph representation where atoms are represented as nodes and bonds as edges. By iteratively passing information contained in features between the nodes, the model builds up atom features, that are finally pooled (summed or averaged) to create a fixed length learned feature vector, often referred to as a molecular fingerprint, that can be passed through a feed-forward network. We also tested a variant of an MPNN in which the encoder acts on edges (bonds) instead of nodes (atoms); this architecture is called a directed MPNN (D-MPNN) and has been shown to have state-of-the-art performance for molecular property prediction [18,66].

All models were trained in multi-task mode to predict all parameters. All neural network models (FFN, MPNN and D-MPNN) were trained for 1000 epochs to minimize the mean squared error loss between the predicted and regressed PCP-SAFT parameters using the Adam optimization solver [69] and the Noam scheduler [70]. The best model checkpoint according to validation loss was used. The learning rate was tuned for each model. We found that using dropout after the pooling step in the MPNN and D-MPNN improved generalization performance. For the MPNN and D-MPNN, the sum pooling function was used (cf. [71]).

The hyperparameters for each model were explored using a quasi-random design with a budget of 100 trials [72]. The best hyperparameters for each model architecture were selected by evaluating the sum of the validation RMSE for all PCP-SAFT parameters. All the final hyperparameters can be found in Table S4.

We experimented with two adaptations of ML to PCP-SAFT prediction. First, since we could already distinguish between associating and non-associating molecules using the heuristic from our regression (i.e., checking the number of association sites), we automatically clamped the association parameters ϵ_{AB} and κ_{AB} to zero for non-associating compounds. We evaluated this clamping of non-associating molecules both as a post-processing step for all models and, for the neural networks, inside the loss function of the neural network. Second, we observed that there were more non-associating than associating molecules in the dataset. Therefore, we tested oversampling of associating molecules in each batch during neural network training using a weighted random sampler:

$$w_i^A = \frac{1}{n_A} \tag{3}$$

where w_i^A is weight for a molecule i with association status A and n_A is the number of molecules of that association status in the whole dataset. We call this oversampling procedure balanced association sampling.

2.5. Evaluation of predictive PCP-SAFT methods

To evaluate ML-SAFT models and the baseline predictive PCP-SAFT methods, a set of 79 molecules was held out from training any models and only used for testing. These molecules were selected such that the majority could be predicted by SEPP and also had regressed parameters. We then split the remaining 905 molecules into training and validation (5%) sets using a clustering procedure. Specifically, ECFP fingerprints with 2048 bits were generated using RDKit, and the k-means clustering algorithm [73] was run on five dimensional projections of these fingerprints from UMAP [74]. We found three clusters to most effectively model the data, as shown in Fig. 4b. Upon manual inspection, we found that the clusters represented chemically interpretable classes of molecules such as alkanes and aromatics. Finally, the molecules were assigned to the training and validation sets so that cluster proportions in each split matched the cluster proportions in the overall dataset using the Stratified Shuffle Split method in scikit-learn [75]. This ensured that each split had a balanced set of molecules. As shown in Fig. 4c, the functional groups in the train and validation splits were balanced, and the test set had even numbers of molecules in each cluster.

We used two metrics for evaluation of the models. For the evaluation of the error between the parameter predictions and regressed parameters, we applied the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}} \tag{4}$$

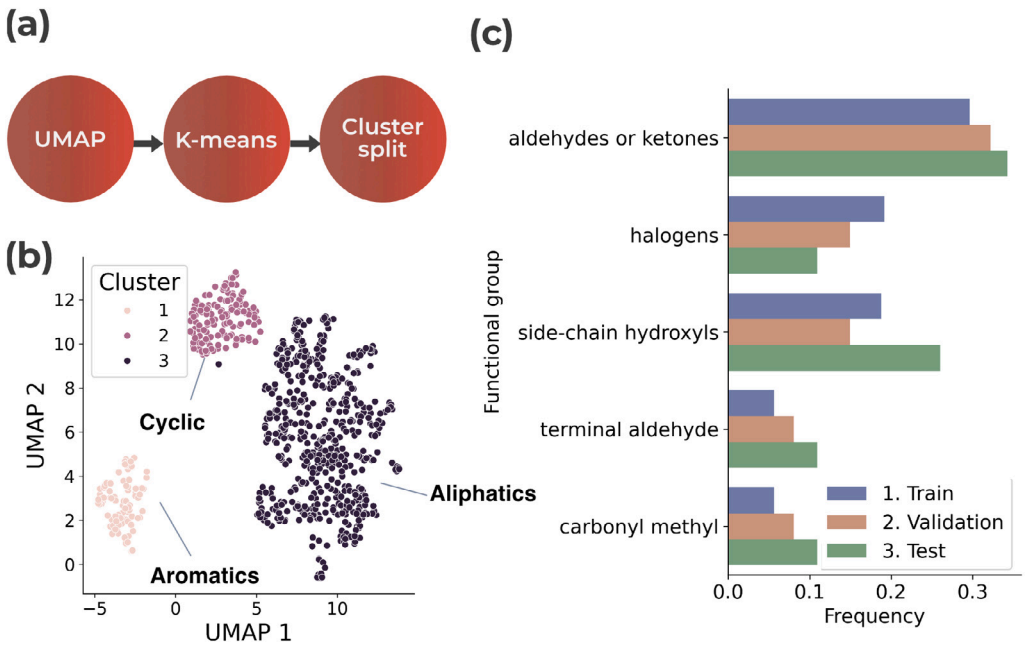

Fig. 4. Data splitting for ML-SAFT datasets. (a) Schematic of the workflow for stratified splitting of the ML-SAFT dataset. UMAP [74] is used for dimensionality reduction of 2048 bit ECFP fingerprints followed by k-means clustering [73] and cluster splitting using stratified shuffle split in scikit-learn [75]. (b) 2D visualization of the clustering using UMAP. (c) The frequency of the top five functional groups in each split are shown. The different functional groups are well-balanced between splits.

where y_i is the regressed PCP-SAFT parameter and $\hat{y}_i$ is the predicted PCP-SAFT parameter. For evaluation of the predictions of density and vapor pressure, we used the percent absolute average deviation (% AAD):

$$\%AAD = \left| \sum_j \frac{Q_J - \hat{Q}_j}{Q_j} \right| * 100 \tag{5}$$

where Q is the experimental value of vapor pressure or liquid density and $\hat{Q}$ is the corresponding PCP-SAFT prediction. Note that, to our knowledge, % AAD is the same as AARD used in some other publications [76].

3. Results

3.1. A robust regression method for PCP-SAFT parameters

We sought to develop an automated approach to regressing the PCP-SAFT parameters from experimental data. Since we used the same initial guess for the regression of all 871 molecules in our dataset, we first aimed to understand the quality of this initial guess across the dataset. As shown in Fig. 5(a–b), the standard initial guess gave liquid density initialization with 35%AAD on average, while the initial accuracy for vapor pressure predictions were significantly worse with an average of 367%AAD. The larger errors for vapor pressure are likely due to the values for vapor pressure varying over several orders of magnitude. However, after regression, most of the PCP-SAFT predictions had less than 5%AAD, and the overall average was 2.31%AAD for vapor pressure predictions and 0.33%AAD for liquid density predictions, as shown in Fig. 5(c–d).

We compared our regressed parameters to a subset of molecules that also had parameters regressed in the literature (see Figure S3 and S4). There were discrepancies between our regressed parameters and the literature values, particularly for associating molecules, reinforcing that parameter degeneracy is a challenge for PCP-SAFT regression [38]. As we show in Figure S10 and S11, after models are trained on the regressed data, model uncertainty could potentially be used to indicate cases where parameter degeneracy is taking place; we find that a large uncertainty in the model predictions is correlated with uncertainty in the parameters regressed from experimental data.

Table 3
RMSE (lower is better) of each model architecture. The best score for each target is marked in bold. RF: Random forest, FFN: Feed-forward neural network, MPNN: Message-passing neural network, D-MPNN: Directed message-passing neural network.

	RF	FFN	MPNN	D-MPNN
m	**0.44**	0.72	0.88	0.74
σ	**0.14**	0.24	0.29	0.29
ϵ/k	**16.67**	32.85	30.79	37.68
ϵ_{AB}	**172.25**	315.24	450.23	376.38

3.2. ML-SAFT accurately predicts regressed PCP-SAFT parameters

To evaluate the accuracy of ML models trained to predict the regressed PCP-SAFT parameters, we first compared the PCP-SAFT parameter predictions from the ML models with the regressed PCP-SAFT parameters. Table 3 shows the RMSE of PCP-SAFT parameter predictions from the various machine learning architectures (full parity plots are shown in the SI). The RF with ECFP fingerprints overall performed best in predicting PCP-SAFT parameters. The FFN, MPNN and D-MPNN models perform worse, likely because they require more data for better performance [77]. However, it most important to look at the end thermodynamic prediction accuracy.

Table 4 presents the absolute average deviation from experimental data of PCP-SAFT predictions of vapor pressure and liquid density using the predicted PCP-SAFT parameters from various ML models. The RF model gave the most accurate predictions for the vapor pressure with of 40% AAD, while the MPNN gave more accurate predictions the liquid density for 8.3% AAD for the molecules in the test set. For the best RF model, the error in vapor pressure and density comes primarily from error in m. As shown in the PCP-SAFT parameter parity plots (Figure S6), the RF consistently underestimates large m. As an example molecule, 1-Perfluoroethyldecalosin has the highest overall error of any molecule in the test set for vapor pressure predictions using the parameters from the Random Forest (see Fig. 6). Here, the error in the vapor pressure measurement is primarily driven by the large error in m, as shown in Table 5. Similarly, to look at an associating molecule, the errors for 3-heptanol are primarily due to errors in m as shown in Table 6.

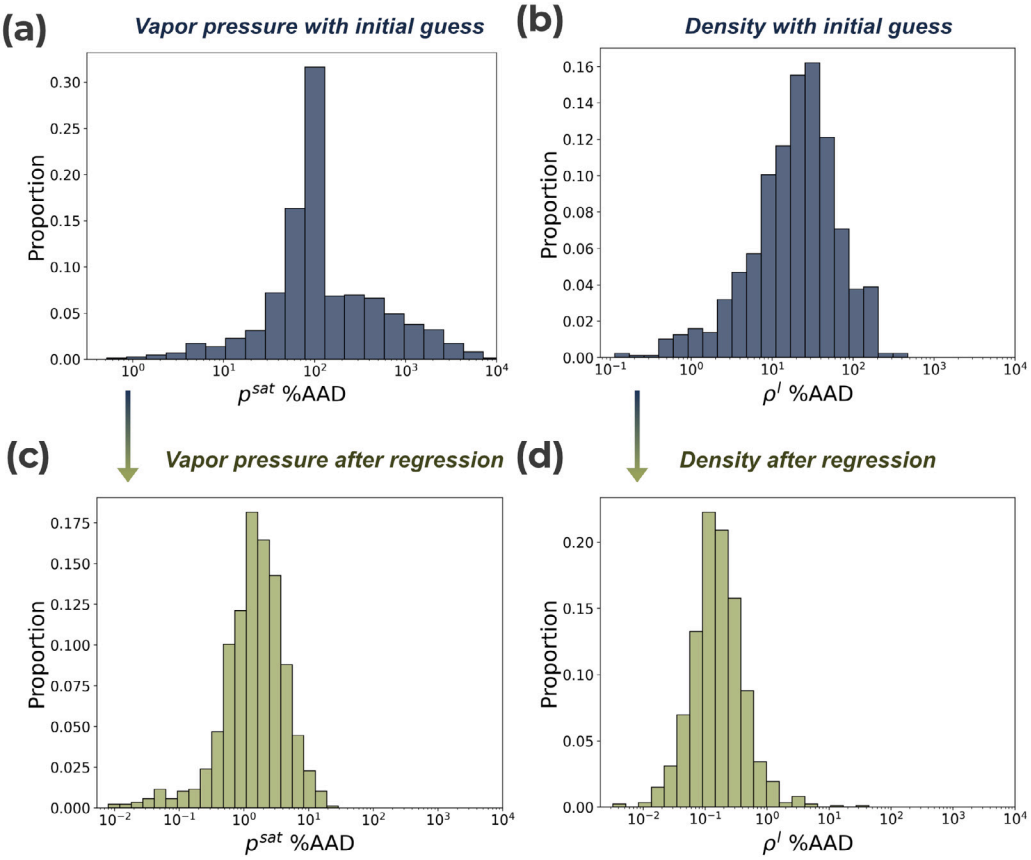

Fig. 5. Distribution of %AAD when using PCP-SAFT regressed parameters for all molecules in the ML-SAFT dataset. (a) Initial guess vapor pressure (b) Initial guess liquid density (c) Regressed vapor pressure (d) Regressed liquid density.

Table 4
Comparison of thermodynamic predictions using PCP-SAFT parameters predicted by ML-SAFT models only. The best score for each thermodynamic quantity is marked in bold. n is the number of molecules in the test set that each method can predict.

	FFN	MPNN	D-MPNN	RF	Regressed
n	73	73	73	73	73
%AAD p_{sat}	1403.93	79.32	61.46	**39.93**	2.46
%AAD ρ^{L}	112.25	11.90	8.67	**8.32**	0.28

Table 5
PCP-SAFT parameters for 1-Perfluoroethyldecalosin.

	RF	Regressed
m	4.64	6.46
σ	3.75	3.78
ϵ/k	191	200
ϵ_{AB}	0	0
κ_{AB}	0	0

Table 6
PCP-SAFT parameters for 3-Heptanol.

	RF	Regressed
m	3.91	4.27
σ	3.60	3.58
ϵ/k	241	241
ϵ_{AB}	1838	2054
κ_{AB}	0.01	0.01

Additional evaluations for models trained via cross-validation yield similar results and can be found in Table S6; these results are consistent with the results in Table 4. Additionally, in SI section S.9, we show that there is not a strong correlation of error in vapor pressure or density prediction with temperature or pressure.

We also note that we experimented with several methods to adapt neural network training to PCP-SAFT parameter prediction. In our experiments, we found that there was no significant difference between clamping the values of the association parameters to zero as a post-processing step versus during training. Furthermore, balanced association sampling did not offer any noticeable improvement in the accuracy of PCP-SAFT parameter predictions (see Table S4). While balanced association sampling did improve predictions of the association parameter ϵ_{AB}, it also degraded the prediction accuracy of the other PCP-SAFT parameters and ultimately led to worse performance on the thermodynamic predictions. Full results of hyperparameter tuning can be found in Table S4.

To answer the question of whether certain classes of compounds have high errors for vapor pressure and density predictions, we plotted in Figs. 7 and 8 the distribution of % AAD for various functional groups. Generally, errors for most functional groups are similar with a small number of outliers in alkanes and halogens.

3.3. Comparison to existing predictive PCP-SAFT methods

We compared ML-SAFT to predictions from the QM method SEPP [14] and the group contribution from Sauer et al. [46]. Please note that the number of test molecules reduces to 65 as SEPP could not provide predictions to 9 molecules due to its inability to predict halogens. When comparing to SEPP, the RF produces more accurate vapor pressure predictions, while SEPP leads to more accurate density predictions, as

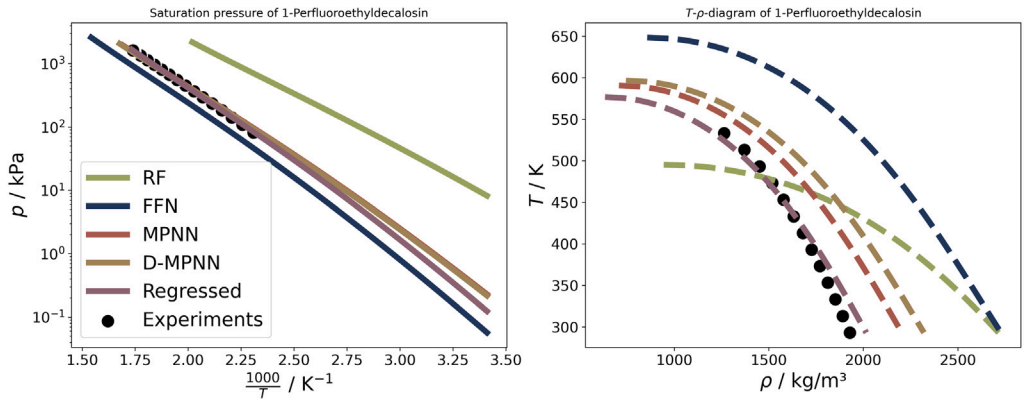

Fig. 6. PCP-SAFT predictions of vapor pressure and density data for 1-Perfluoroethyldecalosin using parameters from various models as well as using the regressed parameters. The dashed line in the density plot represents liquid density.

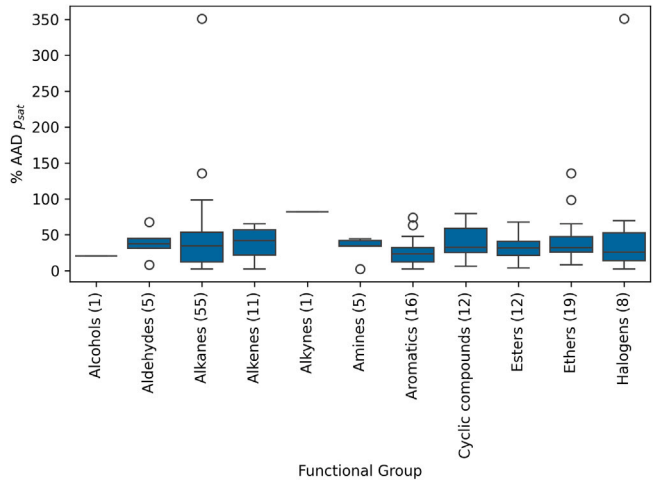

Fig. 7. % AAD in vapor pressure by functional group using PCP-SAFT parameters predicted by the RF. Groups are adapted from Ruggeri and Takahama (see Supplementary Material) [48]. The number of molecules in the test set in each molecular family are shown in parentheses.

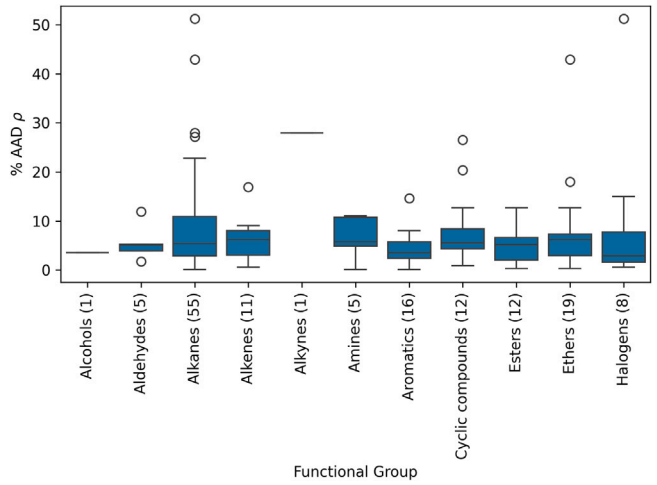

Fig. 8. % AAD in density by functional group using PCP-SAFT parameters predicted by the RF. Groups are adapted from Ruggeri and Takahama (see Supplementary Material) [48]. The number of molecules in the test set in each molecular family are shown in parentheses.

Table 7
Comparison of thermodynamic predictions using PCP-SAFT parameters predicted by ML-SAFT models and SEPP [14]. The best score for each thermodynamic quantity is marked in bold. *n* is the number of molecules in the test set that each method can predict.

	FFN	SEPP	MPNN	D-MPNN	RF	Regressed
n	65	65	65	65	65	65
%AAD p_{sat}	1553.23	107.62	78.30	61.74	**36.50**	2.48
%AAD ρ^L	124.02	**4.49**	11.64	8.77	8.10	0.25

Table 8
Comparison of thermodynamic predictions using PCP-SAFT parameters predicted by ML-SAFT models and a group contribution method (GC) [46]. The best score for each thermodynamic quantity is marked in bold. *n* is the number of molecules in the test set that each method can predict.

	GC	FFN	D-MPNN	MPNN	RF	Regressed
n	11	11	11	11	11	11
%AAD p_{sat}	123.68	83.60	61.43	55.92	**35.28**	1.49
%AAD ρ^L	8.44	29.68	**8.23**	15.66	13.11	0.17

shown in Table 7. However, SEPP has a significant associated computational cost that can extend into days, including conformer generation, two DFT calculations, and a COSMO calculation. In contrast, ML-SAFT methods immediately predict the PCP-SAFT parameters from a SMILES string in milliseconds for each molecule while still maintaining a competitive predictive accuracy.

Comparison with the group contribution method was impaired by the need to convert molecules to groups prior to predictions. Only 11 of the molecules in our test set had functional groups that were already parameterized in the database by Sauer et al. [46]. For this small group of molecules, the RF predictions were significantly more accurate than the GC method for vapor pressure. For the liquid density, the predictions by D-MPNN were more accurate than those of the GC method. Therefore, ML-SAFT resulted in higher accuracy compared to the GC method (see Table 8).

4. Discussion

We proposed ML-SAFT, a machine learning framework for prediction of PCP-SAFT parameters directly from molecular structures. We developed a large database of PCP-SAFT parameters (871 molecules) derived from the Dortmund Data Bank. ML-SAFT models trained on this dataset quickly predicted the regressed PCP-SAFT parameters, and these predicted PCP-SAFT parameters could be in turn used for accurate predictions of thermodynamic quantities. Random forests had the highest accuracy for the thermodynamic predictions of vapor pressure, while D-MPNN also shows promising results for liquid density prediction.

The best ML-SAFT models perform comparably with or better than established/non-ML predictive PCP-SAFT methods, such as SEPP and GC, predictive PCP-SAFT methods while being applicable to a wider range of molecules and giving fast predictions. Group contribution methods require new molecules to be fragmented into groups, and we found that a large fraction of molecules in our dataset were missing parameterized groups or could not be resolved by the automatic fragmentation algorithm. On the other hand, the QM method used for comparison, SEPP, currently is restricted to molecules without halogens as the linear regression model was only fit on alkanes. Furthermore, SEPP requires significant computational time for each molecule, while ML-SAFT affords accurate predictions on a wide range of molecules in milliseconds. We note that for some compound classes, such as amines, data for model training is very limited or not readily available at all, so predictions to these classes should be treated with caution. Here, more data will be required to extend the applicability domain of the models.

There are several ways in which ML-SAFT could be improved. First, we solely used vapor pressure and density data for regression, which can result in parameter degeneracy. Including more experimental data could potentially remove this issue. Second, the training data for ML-SAFT was primarily small molecules with less than 15 atoms. Previous work has shown that PCP-SAFT can effectively predict properties of larger drug-like molecules (e.g., solubility) [78], and the success of RFs and MPNNs in predicting the properties of drug-like molecules suggests that ML-SAFT would be effective given sufficient training data. Please note that drug-like molecules are significantly larger than the molecules used for training and, therefore, predictions on drug-like molecules using the current models are likely not accurate (see Table S7). Third, we do not predict the binary interaction coefficients, which has been shown to significantly improve the quality of PCP-SAFT predictions for mixtures. Future work could address this limitation by training models that contain message-passing between two molecular graphs. This would be a next step towards accurate predictions of multi-component mixture properties using PCP-SAFT.

CRediT authorship contribution statement

Kobi C. Felton: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Lukas Raßpe-Lange:** Methodology, Formal analysis, Conceptualization. **Jan G. Rittig:** Software, Methodology. **Kai Leonhard:** Writing – review & editing, Supervision. **Alexander Mitsos:** Writing – review & editing, Supervision. **Julian Meyer-Kirschner:** Supervision. **Carsten Knösche:** Supervision. **Alexei A. Lapkin:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

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.

Data availability

Code is available; some of the data is made available; most of the original data is under commercial license.

Acknowledgments

K.C.F acknowledges funding from BASF SE, Germany and the Cambridge-Trust Marshall Scholarship. This project was also co-funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany – 466417970 – within the Priority Program “SPP 2331: Machine Learning in Chemical Engineering”. Simulations were performed with computing resources granted by RWTH Aachen University under project “rwth1213”. L.R.L and K.L. gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy Cluster of Excellence 2186, The Fuel Science Center (ID: 390919832). This work was also performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). This work is in part co-funded by the ERDF Project “Innovation Centre in Digital Molecular Technologies”.

Appendix A. Supplementary data 1

Extra figures and hyperparameter tables.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cej.2024.151999.

Appendix B. Supplementary data 2

Code used to produce the results in paper, regressed PCP-SAFT parameters, SMARTS strings used for group contribution identification, scores of predictions from each model, and predicted vapor pressure and density for all molecules in test set.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cej.2024.151999.

References

[1] A. Fredenslund, R.L. Jones, J.M. Prausnitz, Group-contribution estimation of activity coefficients in nonideal liquid mixtures, AiChE J. 21 (6) (1975) 1086–1099, http://dx.doi.org/10.1002/aic.690210607.

[2] S. Skjold-Jorgensen, B. Kolbe, J. Gmehling, P. Rasmussen, Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension, Ind. Eng. Chem. Process Des. Dev. 18 (4) (1979) 714–722, http://dx.doi.org/10.1021/i260072a024.

[3] K. Joback, R. Reid, Estimation of pure component properties from group contributions, Chem. Eng. Commun. 57 (1–6) (1987) 233–243, http://dx.doi.org/10.1080/00986448708960487.

[4] L. Constantinou, R. Gani, New group contribution method for estimating properties of pure compounds, AIChE J. 40 (10) (1994) 1697–1710, http://dx.doi.org/10.1002/aic.690401011.

[5] S. Horstmann, A. Jabłoniec, J. Krafczyk, K. Fischer, J. Gmehling, PSRK group contribution equation of state: comprehensive revision and extension IV, including critical constants and α-function parameters for 1000 components, Fluid Phase Equilib. 227 (2) (2005) 157–164, http://dx.doi.org/10.1016/j.fluid.2004.11.002.

[6] B. Schmid, J. Gmehling, Revised parameters and typical results of the VTPR group contribution equation of state, Fluid Phase Equilib. 317 (2012) 110–126, http://dx.doi.org/10.1016/j.fluid.2012.01.006.

[7] D. Constantinescu, J. Gmehling, Further development of modified UNIFAC (dortmund): Revision and extension 6, J. Chem. Eng. Data 61 (8) (2016) 2738–2748, http://dx.doi.org/10.1021/acs.jced.6b00136.

[8] I. Bell, J. Welliquet, M. Mondejar, A. Bazyleva, S. Quoilin, F. Haglind, Application of the group contribution volume translated peng–robinson equation of state to new commercial refrigerant mixtures, Int. J. Refrig. 103 (2019) 316–328, http://dx.doi.org/10.1016/j.ijrefrig.2019.04.014.

[9] P.J. Walker, A.J. Haslam, A new predictive group-contribution ideal-heat-capacity model and its influence on second-derivative properties calculated using a free-energy equation of state, J. Chem. Eng. Data 65 (12) (2020) 5809–5829, http://dx.doi.org/10.1021/acs.jced.0c00723.

[10] T. Sheldon, B. Giner, C. Adjiman, A. Galindo, G. Jackson, D. Jacquemin, V. Wathelet, E. Perpète, The derivation of size parameters for the SAFT–VR equation of state from quantum mechanical calculations, in: Computer Aided Chemical Engineering, Elsevier, 2006, pp. 143–159, http://dx.doi.org/10.1016/s1570-7946(06)80009-x.

[11] K. Leonhard, N.V. Nhu, K. Lucas, Making equation of state models predictive: Part 2: An improved PCP-SAFT equation of state, Fluid Phase Equilib. 258 (1) (2007) 41–50, http://dx.doi.org/10.1016/j.fluid.2007.05.019.

[12] K. Leonhard, N.V. Nhu, K. Lucas, Making equation of state models predictive-part 3: Improved treatment of multipolar interactions in a PC-SAFT based equation of state, J. Phys. Chem. C 111 (43) (2007) 15533–15543, http://dx.doi.org/10.1021/jp0726081.

[13] R. Fingerhut, W.-L. Chen, A. Schedemann, W. Cordes, J. Rarey, C.-M. Hsieh, J. Vrabec, S.-T. Lin, Comprehensive assessment of COSMO-SAC models for predictions of fluid-phase equilibria, Ind. Eng. Chem. Res. 56 (35) (2017) 9868–9884, http://dx.doi.org/10.1021/acs.iecr.7b01360.

[14] S. Kaminski, K. Leonhard, SEPP: Segment-based equation of state parameter prediction, J. Chem. Eng. Data 65 (12) (2020) 5830–5843, http://dx.doi.org/10.1021/acs.jced.0c00733.

[15] J. Towne, X. Liang, G.M. Kontogeorgis, Application of quantum chemistry insights to the prediction of phase equilibria in associating systems, Ind. Eng. Chem. Res. 60 (16) (2021) 5992–6005, http://dx.doi.org/10.1021/acs.iecr.1c00072.

[16] 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 (3) (2020) 981–985, http://dx.doi.org/10.1021/acs.jpclett.9b03657.

[17] C.A. Grambow, Y.-P. Li, W.H. Green, Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach, J. Phys. Chem. A 123 (27) (2019) 5826–5835, http://dx.doi.org/10.1021/acs.jpca.9b04195.

[18] F.H. Vermeire, W.H. Green, Transfer learning for solvation free energies: From quantum chemistry to experiments, Chem. Eng. J. 418 (2021) 129307, http://dx.doi.org/10.1016/j.cej.2021.129307.

[19] K. Felton, Transfer learning for accelerated process development, 2023, http://dx.doi.org/10.17863/CAM.102031, https://www.repository.cam.ac.uk/handle/1810/358328.

[20] Y. Chung, F.H. Vermeire, H. Wu, P.J. Walker, M.H. Abraham, W.H. Green, Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy, J. Chem. Inf. Model. 62 (3) (2022) 433–446, http://dx.doi.org/10.1021/acs.jcim.1c01103.

[21] E.I.S. Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digit. Discov. 1 (3) (2022) 216–225, http://dx.doi.org/10.1039/d1dd00037c.

[22] S. Qin, S. Jiang, J. Li, P. Balaprakash, R.C.V. Lehn, V.M. Zavala, Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium, Digit. Discov. 2 (1) (2023) 138–151, http://dx.doi.org/10.1039/d2dd00045h.

[23] J.G. Rittig, K. Ben Hicham, A.M. Schweidtmann, M. Dahmen, A. Mitsos, Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids, Comput. Chem. Eng. 171 (2023) 108153, http://dx.doi.org/10.1016/j.compchemeng.2023.108153.

[24] B. Winter, C. Winter, J. Schilling, A. Bardow, A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing, Digit. Discov. (2022) http://dx.doi.org/10.1039/d2dd00058j.

[25] H. Nateghi, G. Sodeifian, F. Razmimanesh, J. Mohebbi Najm Abad, A machine learning approach for thermodynamic modeling of the statically measured solubility of nilotinib hydrochloride monohydrate (anti-cancer drug) in supercritical CO2, Sci. Rep. 13 (1) (2023) http://dx.doi.org/10.1038/s41598-023-40231-4, URL http://dx.doi.org/10.1038/s41598-023-40231-4.

[26] J.G. Rittig, K.C. Felton, A.A. Lapkin, A. Mitsos, Gibbs-duhem-informed neural networks for binary activity coefficient prediction, Digit. Discov. 2 (2023) 1752–1767, http://dx.doi.org/10.1039/D3DD00103B.

[27] F. Abbasi, Z. Abbasi, R.B. Boozarjomehry, Estimation of PC-SAFT binary interaction coefficient by artificial neural network for multicomponent phase equilibrium calculations, Fluid Phase Equilib. 510 (2020) 112486, http://dx.doi.org/10.1016/j.fluid.2020.112486.

[28] H. Matsukawa, M. Kitahara, K. Otake, Estimation of pure component parameters of PC-SAFT EoS by an artificial neural network based on a group contribution method, Fluid Phase Equilib. 548 (2021) 113179, http://dx.doi.org/10.1016/j.fluid.2021.113179.

[29] S.A. Madani, M.-R. Mohammadi, S. Atashrouz, A. Abedi, A. Hemmati-Sarapardeh, A. Mohaddespour, Modeling of nitrogen solubility in normal alkanes using machine learning methods compared with cubic and PC-SAFT equations of state, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-021-03643-8.

[30] A.A. el hadj, M. Laidi, S. Hanini, AI-PCSAFT approach: New high predictive method for estimating PC-SAFT pure component properties and phase equilibria parameters, Fluid Phase Equilib. 555 (2022) 113297, http://dx.doi.org/10.1016/j.fluid.2021.113297.

[31] B. Winter, C. Winter, T. Esper, J. Schilling, A. Bardow, SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients, Fluid Phase Equilib. 568 (2023) 113731, http://dx.doi.org/10.1016/j.fluid.2023.113731.

[32] E.I.S. Medina, S. Linke, M. Stoll, K. Sundmacher, Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution, Digit. Discov. 2 (3) (2023) 781–798, http://dx.doi.org/10.1039/d2dd00142j.

[33] K. Zhu, E.A. Müller, Generating a machine-learned equation of state for fluid properties, J. Phys. Chem. B 124 (39) (2020) 8628–8639, http://dx.doi.org/10.1021/acs.jpcb.0c05806.

[34] D. Rosenberger, K. Barros, T.C. Germann, N. Lubbers, Machine learning of consistent thermodynamic models using automatic differentiation, Phys. Rev. E 105 (4) (2022) http://dx.doi.org/10.1103/physreve.105.045301.

[35] G. Chaparro, E.A. Müller, Development of thermodynamically consistent machine-learning equations of state: Application to the mie fluid, J. Chem. Phys. 158 (18) (2023) http://dx.doi.org/10.1063/5.0146634.

[36] J. Gross, G. Sadowski, Perturbed-chain SAFT: an equation of state based on a perturbation theory for chain molecules, Ind. Eng. Chem. Res. 40 (4) (2001) 1244–1260, http://dx.doi.org/10.1021/ie0003887.

[37] J. Gross, J. Vrabec, An equation-of-state contribution for polar components: Dipolar molecules, AIChE J. 52 (3) (2006) 1194–1204, http://dx.doi.org/10.1002/aic.10683.

[38] J.T. Cripwell, C.E. Schwarz, A.J. Burger, Polar (s)PC-SAFT: Modelling of polar structural isomers and identification of the systematic nature of regression issues, Fluid Phase Equilib. 449 (2017) 156–166, http://dx.doi.org/10.1016/j.fluid.2017.06.027.

[39] J. Habicht, C. Brandenbusch, G. Sadowski, Predicting PC-SAFT pure-component parameters by machine learning using a molecular fingerprint as key input, Fluid Phase Equilib. 565 (2023) 113657, http://dx.doi.org/10.1016/j.fluid.2022.113657.

[40] B. Winter, P. Rehner, T. Esper, J. Schilling, A. Bardow, Understanding the language of molecules: Predicting pure component parameters for the PC-SAFT equation of state from SMILES, 2023, http://dx.doi.org/10.48550/ARXIV.2309.12404.

[41] T. Esper, G. Bauer, P. Rehner, J. Gross, PCP-SAFT parameters of pure substances using large experimental databases, Ind. Eng. Chem. Res. 62 (37) (2023) 15300–15310, http://dx.doi.org/10.1021/acs.iecr.3c02255.

[42] S. Kaminski, Quantum-mechanics-based Prediction of SAFT Parameters for Non-associating and Associating Molecules Containing Carbon, Hydrogen, Oxygen, and Nitrogen (Ph.D. thesis), RWTH Aachen, 2019.

[43] J. Gross, G. Sadowski, Application of the perturbed-chain SAFT equation of state to associating systems, Ind. Eng. Chem. Res. 41 (22) (2002) 5510–5515.

[44] P. Rehner, G. Bauer, J. Gross, Feos: An open-source framework for equations of state and classical density functional theory, Ind. Eng. Chem. Res. (2023) http://dx.doi.org/10.1021/acs.iecr.2c04561.

[45] A. Klamt, Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena, J. Phys. Chem. 99 (7) (1995) 2224–2235, http://dx.doi.org/10.1021/j100007a062, URL http://pubs.acs.org/doi/abs/10.1021/j100007a062.

[46] E. Sauer, M. Stavrou, J. Gross, Comparison between a homo- and a heterosegmented group contribution approach based on the perturbed-chain polar statistical associating fluid theory equation of state, Ind. Eng. Chem. Res. 53 (38) (2014) 14854–14864, http://dx.doi.org/10.1021/ie502203w.

[47] Caleb Bell and Contributors, Thermo: Chemical properties component of chemical engineering design library (ChEDL), 2023, URL https://github.com/CalebBell/thermo.

[48] G. Ruggeri, S. Takahama, Technical note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization, Atmos. Chem. Phys. 16 (7) (2016) 4401–4422, http://dx.doi.org/10.5194/acp-16-4401-2016.

[49] Dortmund databank, 2022, URL www.ddbst.com.

[50] Kobi Felton and Contributors, Pura: Software for cleaning chemical data quickly, 2023, URL https://github.com/sustainable-processes/pura.

[51] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Res. 49 (D1) (2020) D1388–D1395, http://dx.doi.org/10.1093/nar/gkaa971.

[52] NCI/CADD, Chemical identifier resolver, 2023, URL https://cactus.nci.nih.gov/chemical/structure.

[53] D.M. Lowe, P.T. Corbett, P. Murray-Rust, R.C. Glen, Chemical name to structure: OPSIN, an open source solution, J. Chem. Inf. Model. 51 (3) (2011) 739–753, http://dx.doi.org/10.1021/ci100384d.

[54] American Chemical Society, Common chemistry, 2023, URL https://commonchemistry.cas.org.

[55] M. Fitzner, G. Wuitschik, R.J. Koller, J.-M. Adam, T. Schindler, J.-L. Reymond, What can reaction databases teach us about buchwald–hartwig cross-couplings? Chem. Sci. 11 (48) (2020) 13085–13093, http://dx.doi.org/10.1039/d0sc04074f.

[56] H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green, K.F. Jensen, Using machine learning to predict suitable conditions for organic reactions, ACS Cent. Sci. 4 (11) (2018) 1465–1476, http://dx.doi.org/10.1021/acscentsci.8b00357.

[57] G. Landrum, P. Tosco, B. Kelley, Ric, Sriniker, Gedeck, D. Cosgrove, R. Vianello, NadineSchneider, E. Kawashima, D. N, A. Dalke, G. Jones, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, V.F. Scalfani, K. Ujihara, Guillaume Godin, A. Pahl, F. Berenger, JLVarjo, Jasondbiggs, Strets123, JP, Rdkit/rdkit: 2022_09_5 (Q3 2022) release, 2023, http://dx.doi.org/10.5281/zenodo.7671152.

[58] A. de Villiers, C. Schwarz, A. Burger, Improving vapour–liquid-equilibria predictions for mixtures with non-associating polar components using sPC-SAFT extended with two dipolar terms, Fluid Phase Equilib. 305 (2) (2011) 174–184, http://dx.doi.org/10.1016/j.fluid.2011.03.025.

[59] K. Schütt, O. Unke, M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 9377–9388, URL https://proceedings.mlr.press/v139/schutt21a.html.

[60] S. Wang, J. Witek, G.A. Landrum, S. Riniker, Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences, J. Chem. Inf. Model. 60 (4) (2020) 2044–2058, http://dx.doi.org/10.1021/acs.jcim.0c00025.

[61] R. Ramakrishnan, P.O. Dral, M. Rupp, O.A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data 1 (1) (2014) http://dx.doi.org/10.1038/sdata.2014.22.

[62] W.M. Haynes (Ed.), CRC Handbook of Chemistry and Physics, CRC Press, 2014, http://dx.doi.org/10.1201/b17118.

[63] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[64] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754, http://dx.doi.org/10.1021/ci100050t.

[65] B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R.P. Sheridan, V. Pande, Is Multitask Deep Learning Practical for Pharma? J. Chem. Inf. Model. 57 (8) (2017) 2068–2076, http://dx.doi.org/10.1021/acs.jcim.7b00146, URL https://pubs.acs.org/sharingguidelines.

[66] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction, J. Chem. Inf. Model. 59 (8) (2019) 3370–3388, http://dx.doi.org/10.1021/acs.jcim.9b00237, URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237.

[67] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1263–1272, URL https://proceedings.mlr.press/v70/gilmer17a.html.

[68] A.M. Schweidtmann, J.G. Rittig, A. König, M. Grohe, A. Mitsos, M. Dahmen, Graph neural networks for prediction of fuel ignition quality, Energy Fuels 34 (9) (2020) 11395–11407, http://dx.doi.org/10.1021/acs.energyfuels.0c01533.

[69] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, Online. URL http://arxiv.org/abs/1412.6980.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017, Online.

[71] A.M. Schweidtmann, J.G. Rittig, J.M. Weber, M. Grohe, M. Dahmen, K. Leonhard, A. Mitsos, Physical pooling functions in graph neural networks for molecular property prediction, Comput. Chem. Eng. 172 (2023) 108202, http://dx.doi.org/10.1016/j.compchemeng.2023.108202.

[72] O. Bousquet, S. Gelly, K. Kurach, O. Teytaud, D. Vincent, Critical hyper-parameters: No random, no cry, 2017, arXiv:arXiv:1706.03200.

[73] J. MacQueen, Classification and analysis of multivariate observations, in: 5th Berkeley Symp. Math. Statist. Probability, University of California Los Angeles LA USA, 1967, pp. 281–297.

[74] L. McInnes, J. Healy, N. Saul, L. Grossberger, UMAP: Uniform manifold approximation and projection, J. Open. Source Softw. 3 (29) (2018) 861.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[76] E.P. Lyra, L.F.M. Franco, Deriving force fields with a multiscale approach: From ab initio calculations to molecular-based equations of state, J. Chem. Phys. 157 (11) (2022) http://dx.doi.org/10.1063/5.0109350.

[77] E. Heid, C.J. McGill, F.H. Vermeire, W.H. Green, Characterizing uncertainty in machine learning for chemistry, J. Chem. Inf. Model. 63 (13) (2023) 4012–4029, http://dx.doi.org/10.1021/acs.jcim.3c00373.

[78] M. Klajmon, Investigating various parametrization strategies for pharmaceuticals within the PC-SAFT equation of state, J. Chem. Eng. Data 65 (12) (2020) 5753–5767, http://dx.doi.org/10.1021/acs.jced.0c00707.