Graphical Abstract

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

Kobi C. Felton, Lukas Raßpe-Lange, Jan G. Rittig, Kai Leonhard, Alexander Mitsos, Julian Meyer-Kirschner, Carsten Knösche, Alexei A. Lapkin

Highlights

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

Kobi C. Felton, Lukas Raßpe-Lange, Jan G. Rittig, Kai Leonhard, Alexander Mitsos, Julian Meyer-Kirschner, Carsten Knösche, Alexei A. Lapkin

- ML-SAFT is a framework for fast and accurate predictions of PCP-SAFT parameters
- We create the largest available database of regressed PCP-SAFT parameters
- Our best machine learning model has a wide applicability domain

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*
[g]*BASF SE, 67056 Ludwigshafen am Rhein, Germany*

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 for predicting parameters of PCP-SAFT. In order to provide data for training machine learning models, we created the largest dataset of regressed PCP-SAFT parameters in the literature. We then conducted extensive evaluation of several machine learning architectures for predicting PCP-SAFT parameters. We found that our best model provided accurate predictions for a wider range of molecules than existing predictive techniques with 39 % average absolute deviation (AAD) in vapor pressure predictions and 9 % AAD in density predictions.

Keywords: Deep Learning, PC-SAFT, Thermodynamics; Property

*Corresponding author email: aal35@cam.ac.uk

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] and quantum mechanical (QM) simulations [2, 3], recent work has shown that machine learning (ML) models can be used for predictive thermodynamics. This includes methods for predicting infinite dilution activity coefficients from molecular structures and closely related solvation free energies using matrix completion [4] and graph neural networks [5, 6, 7, 8, 9]. However, the limitation of these works is their lack of thermodynamic consistency that comes with rigorously derived equations of state or their inability to predict multiple thermodynamic properties.

To enable general and thermodynamically consistent predictions, one approach is to predict thermodynamic model parameters [10, 11, 12, 13, 14], which also enables simple use in existing process simulation packages. Given the predicted parameters, the thermodynamic model can in turn be used to predict thermodynamic properties. For instance, Winter et al. [14] 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) [15], an established extension of the original PC-SAFT EoS to include polar molecules [16]. 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 [17].

We introduce ML-SAFT, a framework for creating machine learning models that predict PCP-SAFT parameters, shown conceptually in Figure 1. The PCP-SAFT parameters are physically interpretable, but they must be regressed or predicted for each molecule. To train ML models, we developed, to the best of our knowledge, the largest database (988 molecules) of regressed PCP-SAFT parameters using experimental data. We use a combination of

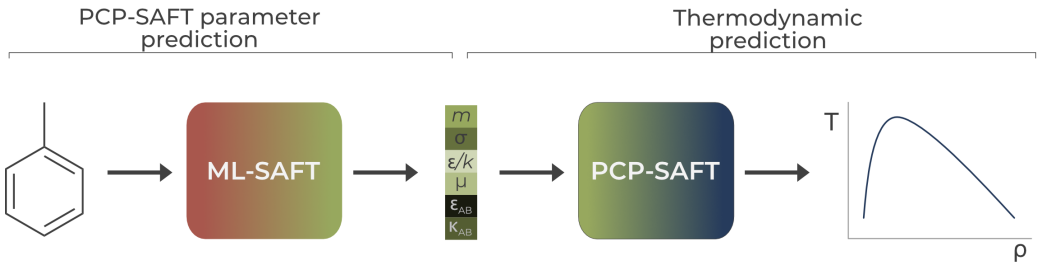

Figure 1: ML-SAFT is a deep learning model for predicting PCP-SAFT parameters directly from molecular structures. PCP-SAFT parameters predicted by ML-SAFT can be used in any PCP-SAFT implementation. Shown schematically is a density prediction.

deep learning and heuristics to enable large-scale automated regression of PCP-SAFT parameters. We then carry out an extensive evaluation of several machine learning architectures for predicting these regressed PCP-SAFT parameters.

We note that Habicht et al. [18] recently developed a feed forward neural network model to predict PC-SAFT parameters from molecular fingerprints. Our framework includes the polar and associating terms and a larger database of regressed PCP-SAFT parameters. We additionally found that random forests outperformed feedforward networks for PCP-SAFT parameter prediction.

2. Methods

2.1. The PCP-SAFT equation of state

The goal of ML-SAFT is to predict the six pure component parameters of PCP-SAFT: σ, the hard sphere diameter; m, the number of segments in a chain of a component; ϵ/k, the depth of the well; μ, the dipole moment; and ϵ_{AB} and κ_{AB}, the association parameters. We use the implementation of PCP-SAFT in FeO_s [19].

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) [3]. 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. An analysis of the surface charge density from COSMO [20] was utilized to calculate the associating parameters ϵ_{AB} and κ_{AB}. 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 from Sauer et al [21] as implemented in FeO_s [19]. For compounds that do not already have groups identified by Sauer et al., we used the group identification from the python package thermo [22] with a modified version of the SMARTS strings from Ruggeri and Takahama (see Supplementary Material) [23].

2.3. Building a dataset for ML-SAFT

2.3.1. Data extraction from the Dortmund Data Bank

Experimental data were extracted from the 2022 Dortmund Data Bank, which contains data for over 40k unique molecules [24]. The software package Pura [25] 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 [26], the Chemical Identifier Resolver [27], OPSIN [28], and the Chemical Abstracts Service [29] 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.

The experimental data was 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 of the range 200-1000 K and pressures outside the range 10-10000 kPa. Densities greater than 2000 kg/m^3 were also excluded. 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

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
κ_{AB}	$0.0 \leq \kappa_{AB} \leq 0.01$	0.0

step has been noted in other attempts to build models on data available in literature databases [30, 31].

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 Figure 2. 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) [3]. The following equation was applied to calculate the sum of squared errors $\mathcal{L}_i$ for molecule i:

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

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} was additionally regressed for associating molecules, while μ and κ_{AB} were not regressed. Instead, we predicted the dipole moment μ and used heuristics to determine if a molecule was associating (described below). The choice to predict the dipole moment is justified for two reasons. First, dipole moment can usually be measured or calculated on a physical basis, and second, previous work has shown that adjusting the dipole moment causes regression to fail due to high correlation with ϵ/k [17, 32].

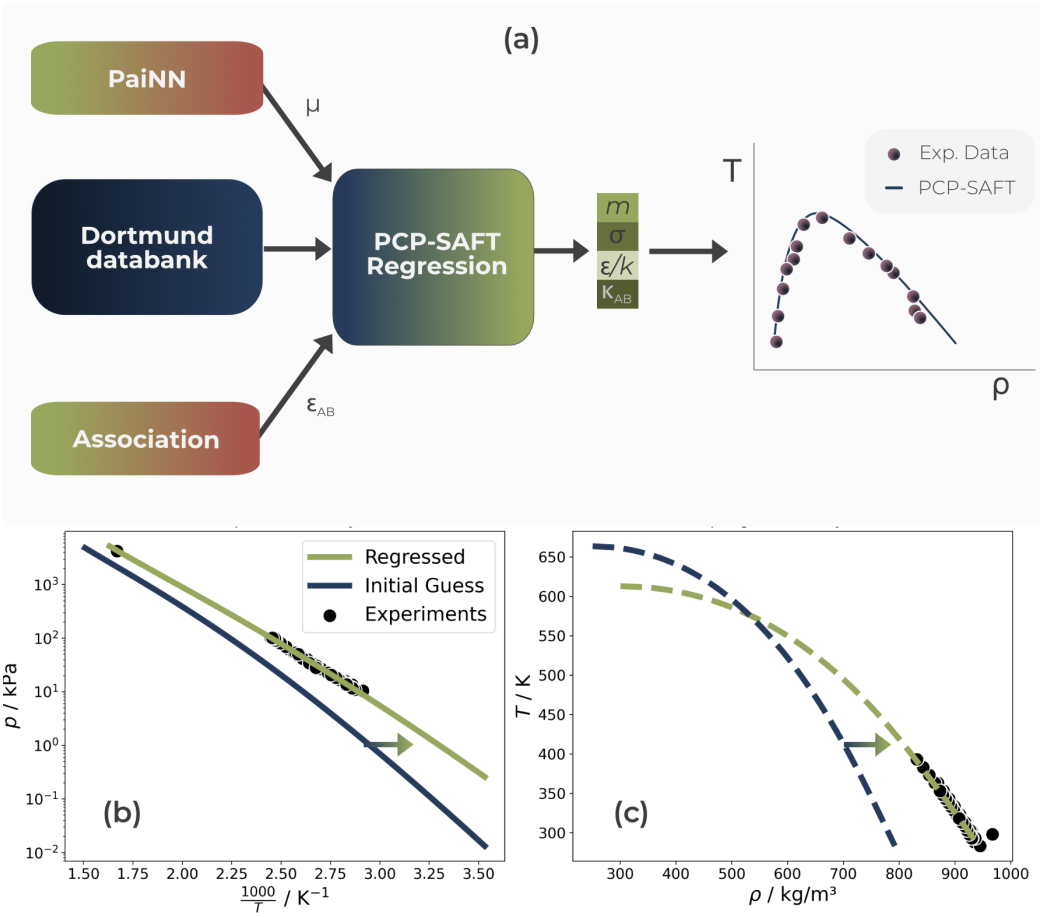

Figure 2: Building a dataset for ML-SAFT: (a) A workflow was developed to automatically regress PCP-SAFT parameters to pure component experimental data. A machine learning model (PaiNN) trained on a combination of DFT and experimental data was used to predict the dipole moments of the experimental dataset, 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-Marquandt algorithm. The dashed line in the density plot represents liquid density.

Therefore, 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 the tensorial equivariant message passing neural network PaiNN developed by Schütt et al. since it has been shown to give accurate predictions of dipole moment [33]. 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 [34]. Subsequently, the dipole moment was calculated using the final vector and tensorial representations of the network:

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

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.

We created two heuristics for improving regression of association parameters. First, non-associating molecules were defined as molecules not containing at least one hydrogen-bond acceptor and donor site via RDKit [35]. The associating parameters ϵ_{AB} and κ_{AB} were set to zero for these non-associating molecules. Second, we found that associating parameter κ_{AB} could be set to 0.01 and not regressed for all associating molecules while maintaining low regression error. With the deep learning predictions of μ and the heuristics for association in place, we successfully regressed PCP-SAFT parameters for the 988 available molecules.

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) [38] and a standard feed-forward network (FFN) that use ECFP4 fingerprints as input [39]. RFs are known to have strong performance for molecular property prediction in drug discovery but

Table 2: Data sets used to train PaiNN architecture for predicting dipole moments. μ_{source} is 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	134k	[36]
CRC	Exp.	RDKit[34]	482	[37]
SEPP	DFT	DFT	1106	See Section 2.2

are less common in process systems engineering [40, 41]. Feed-forward networks were used successfully by Habicht et al. in previous work on predicting PCP-SAFT parameters [18]. Furthermore, we developed a standard message passing neural network (MPNN) [42] that has previously been used to predict several thermodynamic parameters including fuel properties [43] and activity coefficients [7, 9]. 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 [41, 5].

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 optimizer Adam [44] and the Noam scheduler [45]. 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. All the final hyperparameters can be found in Table S1.

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 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 81 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 [46] was run on five dimensional projections of these fingerprints from UMAP [47]. We found three clusters to most effectively model the data, as shown in Figure 3a. 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 [48]. This ensured that each split had a balanced set of molecules. As shown in Figure 3c, the functional groups in the train and validation splits were balanced.

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):

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

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):

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

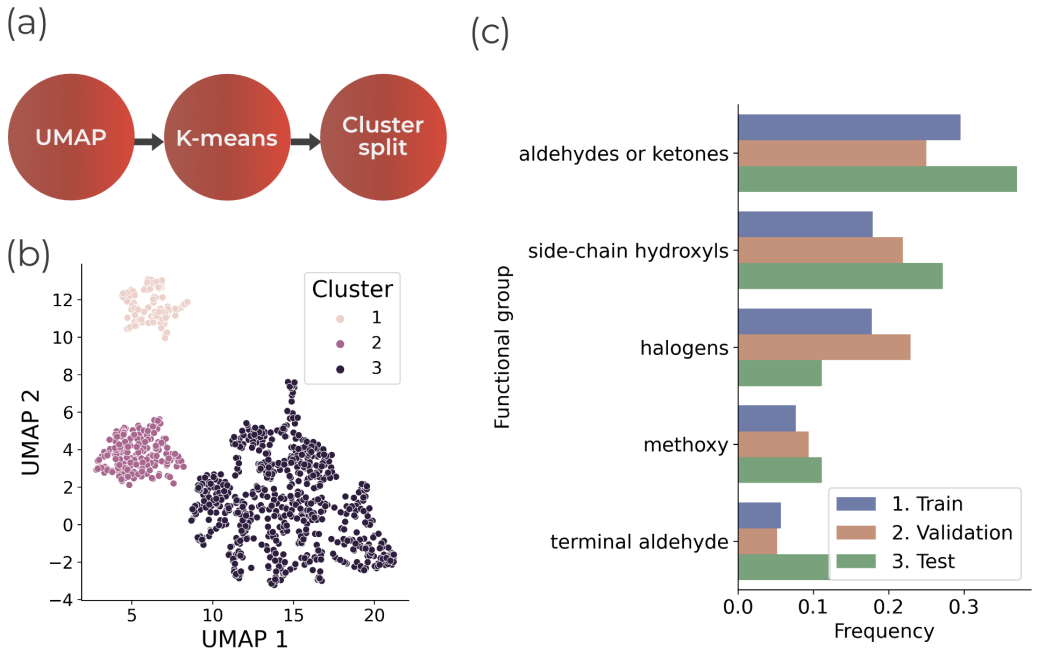

Figure 3: Data splitting for ML-SAFT datasets. (a) Schematic of the workflow for stratified splitting of the ML-SAFT dataset. UMAP [47] is used for dimensionality reduction of 2048 bit ECFP fingerprints followed by k-means clustering [46] and cluster splitting using stratified shuffle split in scikit-learn [48]. (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 Q is the experimental value of vapor pressure or liquid density and $\hat{Q}$ is the corresponding PC-SAFT prediction.

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 988 molecules in our dataset, we first aimed to understand the quality of this initial guess across the dataset. As shown in Figure 4(a-b), the standard initial guess gave liquid density initialization with 35.9 %AAD on average, while the initial accuracy for vapor pressure predictions were significantly worse with an average of 449 %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 using the ML generated PCP-SAFT parameters had less than 5 %AAD, and the overall average was 4.26 %AAD for vapor pressure predictions and 0.62 %AAD for liquid density predictions, as shown in Figure 4(c-d). Empirically, we found that the most important factor for successful regression was the choice of parameter constraints, which we obtained using the maximum and minimum values from all SEPP calculations.

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 RF with ECFP fingerprints performed best in predicting PCP-SAFT parameters. Even after hyperparameter tuning, all neural network architectures had up to 200% worse RMSE values. Compared to previous work by Habicht who found that feed forward neural networks gave accurate predictions, our dataset provides a more difficult regression task as we consider a wider range of molecules and predict polar parameters for associating molecules; this might explain the lower accuracy of the feed forward neural networks in our case. However, the accuracy of predictions of regressed PCP-SAFT parameters might not always translate to the accuracy of thermodynamic predictions, so we also sought to compare the quality of vapor pressure and density predictions.

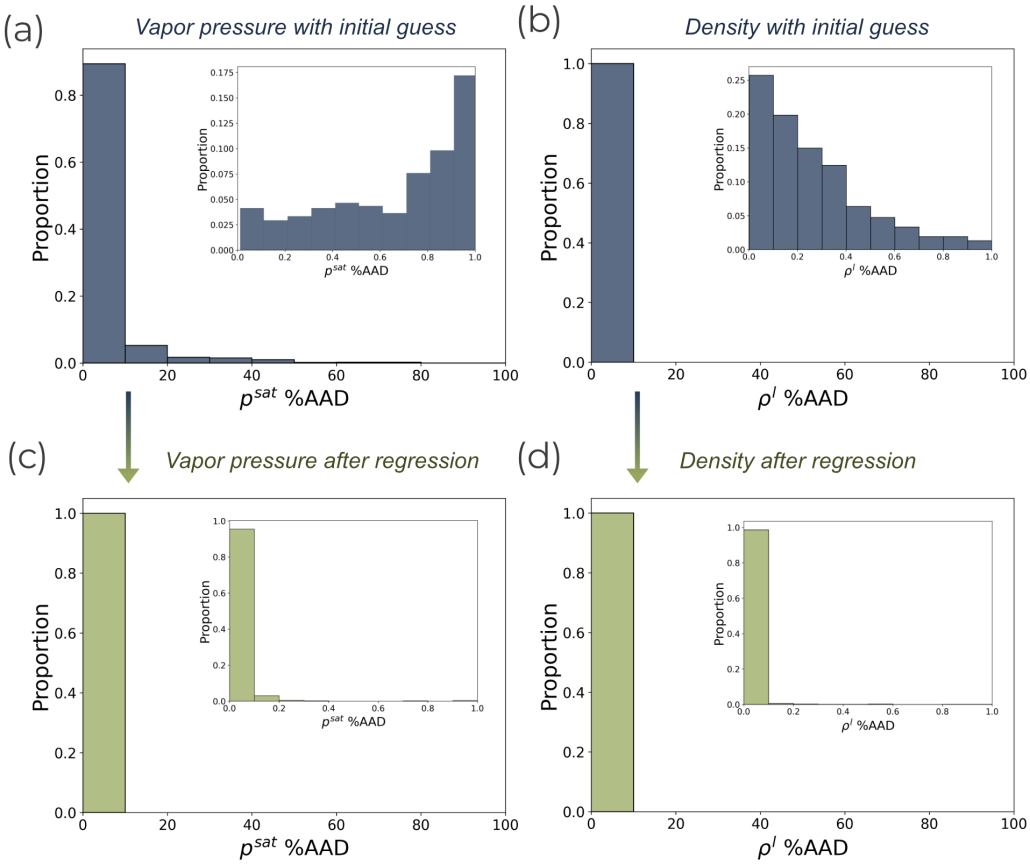

Figure 4: Distribution of %AAD when using PCP-SAFT regressed parameters for all molecules in the ML-SAFT dataset. (a) Initial guess vapor pressure (b) Initial guess liquid density (c) Regressed vapor pressure (d) Regressed liquid density.

Table 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.

	FFN	D-MPNN	MPNN	RF
m	0.59	0.85	0.98	**0.54**
σ	0.28	0.34	0.35	**0.26**
ϵ/k	39.3	39.3	42.7	**31.3**
ϵ_{AB}	362	476	478	**215**

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	D-MPNN	MPNN	RF	Regressed
n	81	81	81	81	81
%AAD p_{sat}	354	53.0	69.7	**38.6**	4.47
%AAD ρ^L	11.3	9.76	13.9	**8.64**	0.800

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 both the vapor pressure and the liquid density with an average of 39% and 9% AAD, respectively, for the molecules in the test set.

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. Although balanced association sampling improved predictions of the association parameter ϵ_{AB}, it 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 S1.

3.3. Comparison to existing predictive PCP-SAFT methods

We compared ML-SAFT to predictions from the QM method SEPP [3] and the group contribution from Sauer et al [21]. Please note that the number of test molecules reduces to 72 as SEPP could not provide predictions to 11 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 shown in Table 5. However, SEPP has a significant associated computational cost that can extend into days, including conformer generation, two DFT calculations, and

Table 5: Comparison of thermodynamic predictions using PCP-SAFT parameters predicted by ML-SAFT models and SEPP [3]. 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	D-MPNN	MPNN	RF	SEPP	Regressed
n	72	72	72	72	72	72
%AAD p_{sat}	372	49.2	67.8	**39.5**	111	4.51
%AAD ρ^L	11.7	9.54	13.3	8.88	**5.10**	0.82

Table 6: Comparison of thermodynamic predictions using PCP-SAFT parameters predicted by ML-SAFT models and a group contribution method (GC) [21]. 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	D-MPNN	MPNN	RF	GC	Regressed
n	13	13	13	13	13	13
%AAD p_{sat}	132	41.1	65.7	**39.6**	118	3.28
%AAD ρ^L	13.2	**9.98**	15.8	12.2	10.8	2.98

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 13 of the molecules in our test set had functional groups that were already parametrized in the database by Sauer et al [21]. For this small group of molecules, the RF predictions were significantly more accurate than the GC method for vapor pressure, while the D-MPNN predictions performed best for density.

4. Discussion

We proposed ML-SAFT, a machine learning framework for prediction of PCP-SAFT parameters directly from molecular structures. We developed the largest database of PCP-SAFT parameters (988 molecules) derived from the Dortmund databank. ML-SAFT trained on this dataset accurately pre-

dicted 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 regressed PCP-SAFT parameters and the thermodynamic predictions overall.

The best ML-SAFT model (random forests) performs comparably with or better than existing 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.

There are several ways in which ML-SAFT could be improved. First, 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) [49], and the success of MPNNs in predicting the properties of drug-like molecules suggests that ML-SAFT would be effective given sufficient training data. Second, 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.

5. Author Contributions

K.C.F developed the concept of ML-SAFT, created the dataset, trained the dipole moment prediction model and all ML-SAFT parameter prediction models and wrote the majority of the manuscript. L.R. executed the SEPP calculations, assisted in the PCP-SAFT parameter regression and assisted in manuscript preparation. J.G.R. implemented the MPNN and dipole moment prediction model and assisted in manuscript preparation. K.L., A.M., J.M.-K., C.K, and A.A.L. acquired funding, provided supervision and edited the manuscript.

6. Acknowledgements

K.C.F acknowledges funding from BASF SE and the Cambridge-Trust Marshall Scholarship. This project was also co-funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466417970 – within the Priority Programme "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 ResearchFoundation) 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.

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.

References

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

[2] 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, Industrial and Engineering Chemistry Research 56 (35) (2017) 9868–9884. doi:10.1021/acs.iecr.7b01360.
URL https://doi.org/10.1021/acs.iecr.7b01360

[3] S. Kaminski, K. Leonhard, SEPP: Segment-based equation of state parameter prediction, Journal of Chemical and Engineering Data 65 (12) (2020) 5830–5843. doi:10.1021/acs.jced.0c00733.
URL https://doi.org/10.1021/acs.jced.0c00733

[4] 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, The Journal of Physical Chemistry Letters 11 (3) (2020) 981–985. doi:10.1021/acs.jpclett.9b03657.
URL https://doi.org/10.1021/acs.jpclett.9b03657

[5] F. H. Vermeire, W. H. Green, Transfer learning for solvation free energies: From quantum chemistry to experiments, Chemical Engineering Journal 418 (2021) 129307. doi:10.1016/j. cej.2021.129307.
URL https://doi.org/10.1016/j .cej.2021.129307

[6] K. C. Felton, H. Ben-Safar, A. Lapkin, Deepgamma: A deep learning model for activity coefficient prediction (2022).

[7] E. I. S. Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digital Discovery 1 (3) (2022) 216–225. doi:10.1039/d1dd00037c.
URL https://doi.org/10.1039/d1dd00037c

[8] 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, Digital Discovery 2 (1) (2023) 138–151. doi:10.1039/d2dd00045h.
URL https://doi.org/10.1039/d2dd00045h

[9] 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, Computers and Chemical Engineering 171 (2023) 108153. doi:10.1016/j.compchemeng.2023.108153.
URL https://doi.org/10.1016/j.compchemeng.2023.108153

[10] 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 Equilibria 510 (2020) 112486. doi:10.1016/j.fluid.2020.112486.
URL https://doi.org/10.1016/j.fluid.2020.112486

[11] 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 Equilibria 548 (2021) 113179. doi:10.1016/j.fluid.2021.113179.
URL https://doi.org/10.1016/j.fluid.2021.113179

[12] 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, Scientific Reports 11 (1) (Dec. 2021). doi:10.1038/s41598-021-03643-8.
URL https://doi.org/10.1038/s41598-021-03643-8

[13] 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 Equilibria 555 (2022) 113297. doi:10.1016/j.fluid.2021.113297.
URL https://doi.org/10.1016/j.fluid.2021.113297

[14] B. Winter, C. Winter, J. Schilling, A. Bardow, A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing, Digital Discovery (2022). doi:10.1039/d2dd00058j.
URL https://doi.org/10.1039/d2dd00058j

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

[16] J. Gross, G. Sadowski, Perturbed-chain SAFT: an equation of state based on a perturbation theory for chain molecules, Industrial and Engineering Chemistry Research 40 (4) (2001) 1244–1260. doi:10.1021/ie0003887.
URL https://doi.org/10.1021/ie0003887

[17] 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 Equilibria 449 (2017) 156–166. doi:10.1016/j.fluid.2017.06.027.
URL https://doi.org/10.1016/j.fluid.2017.06.027

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

[19] P. Rehner, G. Bauer, J. Gross, Feos: An open-source framework for equations of state and classical density functional theory, Industrial and Engineering Chemistry Research (2023). doi:10.1021/acs.iecr.2c04561.
URL https://doi.org/10.1021/acs.iecr.2c04561

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

[21] 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, Industrial and Engineering Chemistry Research 53 (38) (2014) 14854–14864. doi:10.1021/ie502203w.
URL https://doi.org/10.1021/ie502203w

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

[23] G. Ruggeri, S. Takahama, Technical note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization, Atmospheric Chemistry and Physics 16 (7) (2016) 4401–4422. doi:10.5194/acp-16-4401-2016.
URL https://doi.org/10.5194/2Facp-16-4401-2016

[24] Dortmund databank (2022).
URL www.ddbst.com

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

[26] 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 Research 49 (D1) (2020) D1388–D1395. doi:10.1093/nar/gkaa971.

[27] NCI/CADD, Chemical identifier resolver.
URL cactus.nci.nih.gov/chemical/structure

[28] D. M. Lowe, P. T. Corbett, P. Murray-Rust, R. C. Glen, Chemical name to structure: OPSIN, an open source solution, Journal of Chemical Information and Modeling 51 (3) (2011) 739–753. doi:10.1021/ci100384d.
URL https://doi.org/10.1021/ci100384d

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

[30] 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?, Chemical Science 11 (48) (2020) 13085–13093. doi:10.1039/d0 sc04074f.
URL https://doi.org/10.1039/d0 sc04074f

[31] 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 Central Science 4 (11) (2018) 1465–1476. doi:10.1021/acscentsci.8b00357.
URL https://doi.org/10.1021/acscentsci.8b00357

[32] 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 Equilibria 305 (2) (2011) 174–184. doi:10.1016/j.fluid.2011.03.025.
URL https://doi.org/10.1016/j.fluid.2011.03.025

[33] 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, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 9377–9388.
URL https://proceedings.mlr.press/v139/schutt21a.html

[34] 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, Journal of Chemical Information and Modeling 60 (4) (2020) 2044–2058. doi:10.1021/acs.jcim.0c00025.
URL https://doi.org/10.1021/acs.jcim.0c00025

[35] 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 (Feb. 2023). doi:10.5281/zenodo.7671152.
URL https://doi.org/10.5281/zenodo.7671152

[36] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data 1 (1) (Aug. 2014). doi:10.1038/sdata.2014.22.
URL https://doi.org/10.1038/sdata.2014.22

[37] W. M. Haynes (Ed.), CRC Handbook of Chemistry and Physics, CRC

Press, 2014. doi:10.1201/b17118.
URL https://doi.org/10.1201/b17118

[38] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[39] D. Rogers, M. Hahn, Extended-connectivity fingerprints, Journal of Chemical Information and Modeling 50 (5) (2010) 742–754. doi:10.1021/ci100050t.
URL https://doi.org/10.1021/ci100050t

[40] 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. doi:10.1021/acs.jcim.7b00146.
URL https://pubs.acs.org/sharingguidelines

[41] 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. doi:10.1021/acs.jcim.9b00237.
URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237

[42] 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, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1263–1272.
URL https://proceedings.mlr.press/v70/gilmer17a.html

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

[44] D. P. Kingma, J. Ba, Y. Bengio, Y. LeCun, 3rd international conference on learning representations, ICLR, San Diego (2015).

[45] 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, Vol. 30, Curran Associates, Inc., 2017.

[46] 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.

[47] L. McInnes, J. Healy, N. Saul, L. Grossberger, Umap: Uniform manifold approximation and projection, The Journal of Open Source Software 3 (29) (2018) 861.

[48] 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, Journal of Machine Learning Research 12 (2011) 2825–2830.

[49] M. Klajmon, Investigating various parametrization strategies for pharmaceuticals within the PC-SAFT equation of state, Journal of Chemical and Engineering Data 65 (12) (2020) 5753–5767. doi:10.1021/acs.jced.0c00707.
URL https://doi.org/10.1021/acs.jced.0c00707