Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures

Maarten R. Dobbelaere[1], Yannick Ureel[1], Florence H. Vermeire[1], Lowie Tomme[1], Christian V. Stevens[2], Kevin M. Van Geem[1,*]

[1]Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Technologiepark 125, 9052 Gent, Belgium
[2]SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium

[*] Corresponding author: Kevin.VanGeem@UGent.be, Technologiepark 125, 9052 Gent, Belgium;

Keywords: Machine Learning, Physicochemical Property, Hydrocarbons, Chemical Mixtures, Naphtha

Abstract:

Machine learning has proven effective for predicting properties of pure compounds from molecular structures, but properties of mixtures, in particular oil fractions, are rarely dealt with. At best, the bulk properties are estimated based on pure compound properties, linear mixing rules, and a reconstructed composition of the feedstock. As the detailed composition of such mixtures is rarely well determined and often approximated by lumps, the accuracy of the estimated bulk properties can be improved. In this work, we demonstrate for a naphtha case study our bulk property estimation method. First a detailed PIONA composition is delumped into a molecule-level composition, a machine learning based approach is used to predict properties of those molecules, which are further combined in another deep neural network for the prediction of bulk properties. The latter machine learning models are trained on mixture properties using vectors that represent the mixture. The first vector is a linear combination of the molecular representation vectors, and is the representation of the molecular geometries that make up the mixture. The second vector applies linear mixing rules on boiling temperatures, critical temperatures, liquid densities, and vapor pressures that are predicted with machine learning. The last vector consists of a learned distillation curve. We show that an integrated machine learning approach that starts from the molecular structures in the mixture offers significant improvements in predicting mixture properties over existing approaches applied in industry and academia.

1. Introduction

It is essential to have a good knowledge about the properties of chemicals and their mixtures when designing new materials and processes [1]. Complex mixtures, such as crude and renewable oils, need to be reconstructed to a more detailed composition before their properties can be calculated. A whole spectrum of methods exists to reconstruct the composition of complex mixtures [2]. Yet, with all methods it remains difficult to accurately calculate the mixture properties, such as boiling points, density, and viscosity [3]. The inaccurate calculation of mixture properties affects the molecular reconstruction, which in its turn, limits the accuracy of detailed kinetic models. An accurate prediction of mixture density and viscosity is for example essential for kinetic models and reactor design [4]. These mixture properties are often also used in correlations to predict other properties or as quality parameters of hydrocarbon products, so that the prediction has a direct economic impact.

Next to a representative reconstruction of the feedstock, it is equally important to have a rapid and accurate estimation of the individual property. Several methods exist to correlate a wide range of molecular properties solely using the molecular structure [5]. In the chemical industry, it is common practice to use group contribution methods [6-11]. In such a group contribution method, properties are calculated by adding contributions of functional groups. Despite being a fast approach, the accuracy is limited by the empirical, linear nature of the functional group contributions that only depends on the nearest neighbors. Over the last years, machine learning has emerged as an alternative tool for predicting molecular properties, because of its speed and application range [12]. First, it has especially been applied to predict molecular quantum chemical properties in theoretical chemistry studies [13-15]. In chemical engineering, mostly neural-network-based approaches have been developed for a wide range of properties, such as

but not limited to enthalpies of formation [16-18], solvation energies [19, 20], octane numbers [21, 22], boiling points [23-27], and vapor pressure [25, 26]. Alshehri *et al.* [28] published the most extensive study so far that applies both group contribution and machine learning methods to predict 25 pure compound properties. Their dataset contains around 25000 organic molecules up to 30 heavy atoms that can contain 9 heteroatoms. All models in their work are trained on at least 400 data points. In general, one can state that the predictive performance of data-driven models depends on the quality of the data, the amount of datapoints, and the diversity of the data.

When it comes to estimating state properties of oil fractions, further assumptions are often made. One example is the prediction of the density or specific gravity where the most popular approach assumes that the mixture is an ideal liquid mixture, without excess molar volume. This assumption creates an initial error on the mixture predictions, as it does not hold in reality. In addition, the accuracy of the calculated pure compound properties also plays a role [3]. When there are no experimental density values available, densities of pure compounds can be calculated using an equation of state, group contribution methods, or via correlations [29]. These methods are again limited because either data of similar molecules or accurate (pseudo)critical properties are required. The same difficulties exist for other state properties of molecular mixtures, such as viscosity. For dynamic viscosities of liquid mixtures, the most applied equation is the Grunberg-Nissan equation [30], which links the mean viscosity to the viscosity of the pure components in the mixture. Reasonably accurate results within 10% error margin can be expected for binary mixtures [31]. When investigating oils, the mixtures contain more than two compounds and the properties of the individual compounds are typically unknown. An early attempt to predict the dynamic viscosity of mineral oils linked the property to the statistical distribution of carbon atoms in paraffinic chains, aromatic rings and naphthenic

rings [32]. Lohrenz *et al.* [33] determined the viscosity of reservoir fluids by first making a characterization of the involved compounds in the heavy hydrocarbon mixture. This characterization consists of pseudo-compounds for which then individual gas-phase properties are estimated. Using correlations and mixture rules that start from individual properties and the specific gravity of the mixture, the liquid mixture viscosity could then be calculated with a mean absolute error of 16%. Recent studies have applied modern computational tools such as neural networks for predicting mixture properties [34]. Albahri [35] predicted the specific gravity with two single-layer neural networks that used a range of nine boiling points from the ASTM D86 [36] as input values. Plehiers *et al.* [37] trained a model on naphtha samples using a lumped feedstock with 28 pseudo-components as input to predict points of the distillation curve and mixture properties. Although the aforementioned models are able to predict properties with high accuracy, they are limited in applicability range. With the current methods, new feedstocks have not been validated because they fall outside of the application range. Data-driven models that link the prediction of pure component properties and mixture properties are, however, not yet available [38].

In industrial laboratories, experimental equipment is sometimes available but this is not always the case in academic research groups. However, gas chromatography is typically available. Furthermore, the experimental determination of physical properties for certain complex mixtures is nearly impossible due to the reactiveness of these mixtures. Plastic waste pyrolysis oils, which are increasingly investigated as promising intermediates for the chemical recycling of plastic waste, contain a large percentage of highly-reactive olefins [39]. During experimental analyses, these olefin-rich mixtures can suffer of thermal degradation and the composition change leads to unreliable measurements [40]. A solution is, thus, to create a computational approach that combines pure compound and mixture property predictions.

In this work, we show how physical properties of complex hydrocarbon mixtures are linked with molecular properties of individual compounds and with molecular structures. Many studies are dedicated to the relation between molecular properties and the molecular structure [41], as well as to predicting mixture properties from characterizing features of that mixture [37]. However, an integrated machine learning approach that links mixture properties with the molecular structure and their individual properties is still lacking. For a naphtha case study, we use a detailed paraffins, isoparaffins, olefins, naphthenes, aromatics (PIONA) composition matrix as input for the computational method. This matrix is converted into a molecule-based composition via a rule-based algorithm. We also demonstrate that the neural network-based property prediction tool GauL-HDAD [17], originally developed for thermochemical molecular properties, is able to predict normal boiling temperatures, critical temperatures, critical pressures, acentric factors, liquid densities, and vapor pressures of hydrocarbons with good accuracy. By combining the detailed molecule-based composition, the molecular representations and the molecular properties, molecular mixture representations are generated that serve as input for the deep neural networks. The mixture is represented by two vectors: a geometry-based mixture representation and a property-based mixture representation. The complete workflow of predicting individual and mixture properties is available as open-source software. We report the performance on different naphtha properties, namely boiling point curves, specific gravity, viscosity, and surface tension.

2. Methods

2.1. Datasets

2.1.1. Naphtha Space

All hydrocarbon molecules that are likely to be present in naphtha samples are stored in an unlabeled library. In this study, the naphtha space contains molecules with up to 12 carbon atoms and has been determined based on a huge dataset of fossil and plastic waste derived naphtha feedstock. Within this carbon number range, all *n*-paraffins, all isoparaffins, branched and linear olefins with up to two double bonds, monocyclic naphthenes, and monocyclic aromatics are selected as potential naphtha molecules. A constraint is added to the naphthenes so that their ring is either five-membered or six-membered [42]. All other ring sizes are not included. The isomers of all these molecules with up to 12 carbon atoms are generated with surge, an open-source chemical graph generator [43]. The generated SMILES identifiers of the molecules are canonicalized using RDKit [44] since surge does not take into account aromaticity. Molecules that are physically impossible, such as naphthenes with consecutive double bonds are not present, since surge only outputs chemically feasible compounds. The naphtha space counts about 26k molecules.

2.1.2. Molecular Properties

The chemical property handbooks of Carl L. Yaws [45-47] are used to assign experimental molecular properties to the naphtha molecules. The boiling temperature, critical temperature, liquid density, vapor pressure, critical pressure, and acentric factor are included for a subset of the hydrocarbon library, since for most of the molecules in the naphtha space no experimental

datapoints are available. Table 1 gives an insight in the molecular properties that are used for training molecular property prediction models. Normal boiling points, liquid densities, and vapor pressures are only trained on experimental data. Since less than 100 experimental values are available for critical temperatures, critical pressures, and acentric factors, the training sets of these properties also include calculated datapoints. These values are calculated using the Joback group contribution method [6]. The number of experimental and calculated values used for training of each property is shown in Table 1.

The liquid density d is labeled as experimental, but calculated at 293.15 K (20 °C) using the Daubert-Danner correlation [48] (eq **(1)**), for which the experimentally verified coefficients *A*, *B*, *C*, and *n* are reported by Yaws [46].

$$d = A \cdot B^{\left(1-\frac{T}{C}\right)^{-n}} \tag{1}$$

In a similar way, the vapor pressure P at 100 °F (311 K or 37.8 °C) is calculated with the Antoine equation (eq **(2)**), using experimentally verified coefficients A, B, and C from Yaws and Satyro [47].

$$\log_{10} P = A - \frac{B}{C + T} \tag{2}$$

Table 1: Overview of the training data for the individual compound properties

Property	Unit	Experimental Data	Calculated Data	Minimal Value	Maximal Value
Normal boiling point	*K*	1025	0	261.4	536.6
Critical temperature	*K*	93	985	407.8	723.6
Liquid density	*kg m^{-3}*	1117	0	558.2	921.2
Vapor pressure	*log(kPa)*	1025	0	-1.61	3.59
Acentric factor	-	89	1088	0.182	0.576
Critical pressure	*bar*	97	981	18.2	49.0

2.1.3. Naphtha Samples

The naphtha dataset used in this work consists of 382 curated experimental samples, collected from Pyl *et al.* [49] and Mei *et al.* [50]. All samples have a detailed PIONA composition that is compatible with the naphtha space mentioned above. The bulk properties of the included samples are different depending on the source. The 272 samples from Pyl *et al.* [49] contain the initial boiling point (IBP), 50%-boiling point (BP50), final boiling point (FBP), and the specific gravity at 60 °F (15.5 °C). The 50%-boiling point denotes the temperature at which 50 vol% of the mixture is evaporated [49]. The other 110 samples, from Mei *et al.* [50], do not include IBP and FBP, but the boiling points at 5%, 95%, and between 10% and 90% with a step of 10%. Next to boiling points, the liquid density at 20 °C (293.15 K), dynamic viscosity, and surface tension are given. The bulk properties (density, viscosity, surface tension) reported by Mei *et al.* [50] are not experimental, but calculated with Aspen HYSYS. The boiling points of all 382 samples are determined via the ASTM D86 standard test method [36] and converted to true boiling points via the correlation of Riazi [51].

2.2. Delumping Strategy

In order to accurately predict the properties of naphthas starting from the individual components, it is important to have a reasonable estimate of which molecules make up the naphtha. The input to this algorithm is a molecular-type homologous series (MTHS) matrix [52] and each value in this matrix consists of a lump of one or more molecules. Several approaches have been developed to delump the matrix into a molecule-level composition, in order to calculate mixture properties [53-57]. In this work, we adopt an semiempirical approach, similar to the one from Ranzi *et al.* [58], in which an internal distribution is created for each lump.

For light crude hydrocarbon mixtures, such as naphthas, regularities are found in the distribution of the isomers in the mixture [59, 60]. This means that the distribution of molecules within one lump is more or less equal for different mixtures. The absolute fraction of a molecule in naphtha can thus be calculated by multiplication of the absolute fraction of the lump and the internal fraction of the molecule in that lump. The values for isomers in the internal distribution are found by assuming probabilities that a carbon atom can be methylated or alkylated. Different rules are set up for isoparaffins, naphthenes, and aromatics, which can all be found in Supporting Information.

A molecule gets its weight depending on the substructures that are present in the molecule. The molecule is read into RDKit [44] using its SMILES identifier [61, 62] and is then classified in its PIONA class. Based on its carbon number, the molecule is assigned to the corresponding lump. Substructure matching using SMARTS [63] is performed to assign a value to that molecule. The internal value for an isoparaffinic molecule is found via a single empirical formula, given by eq **(3)**.

$$w = \beta \cdot \alpha_{methyl}^{(n_{methyl} - n_{quat})} \cdot \alpha_{ethyl}^{n_{ethyl}} \cdot \alpha_{quat}^{n_{quat}} \quad (3)$$

In eq **(3)** there are several empirical parameters: α_{methyl} is the weight for a methylation, α_{ethyl} is the weight for an ethylation, and α_{quat} is the weight for a quaternary carbon atom. Based on investigation of experimental samples [59, 60, 64-68], the values are set on $\alpha_{methyl} = 0.3$, $\alpha_{ethyl} = 0.05$ and $\alpha_{quat} = 0.05$, which is similar to the values of Ranzi *et al.* [58] which are respectively 0.28, 0.045, and 0.056. The pre-factor β is set at 0 when alkyl groups with more than 2 carbon atoms are present, at 1 for molecules that contain planes of symmetry (e.g. 3-methylpentane), and at 2 for all other compounds. Correction factors are added for 3-methylhexane and 2-methylheptane to agree with experimental observations [59, 60]. The

exponents n_{methyl}, n_{ethyl}, and n_{quat} are respectively the number of methyl groups, ethyl groups, and quaternary atoms. The eventual internal fraction x of a molecule in a lump is found by dividing the weight w by the sum of the weights of all L molecules in the lump, as shown in eq **(4)**.

$$x = \sum_{i}^{L} w_i \tag{4}$$

Figure 1 illustrates the underlying distribution of the C_8 isoparaffins lump obtained with the fitted values from this work and compared to the distributions from Ranzi *et al.* [56] and Mei *et al.* [57]. For visualization reasons, the dimethylhexanes, ethylmethylpentanes, and trimethylpentanes are grouped. The detailed underlying distribution of this lump can be found in Supporting Information. The samples named Ponca, Occidental, and Texas are real crude oil samples with an experimentally determined composition [56], while the other distributions are generated with delumping rules.

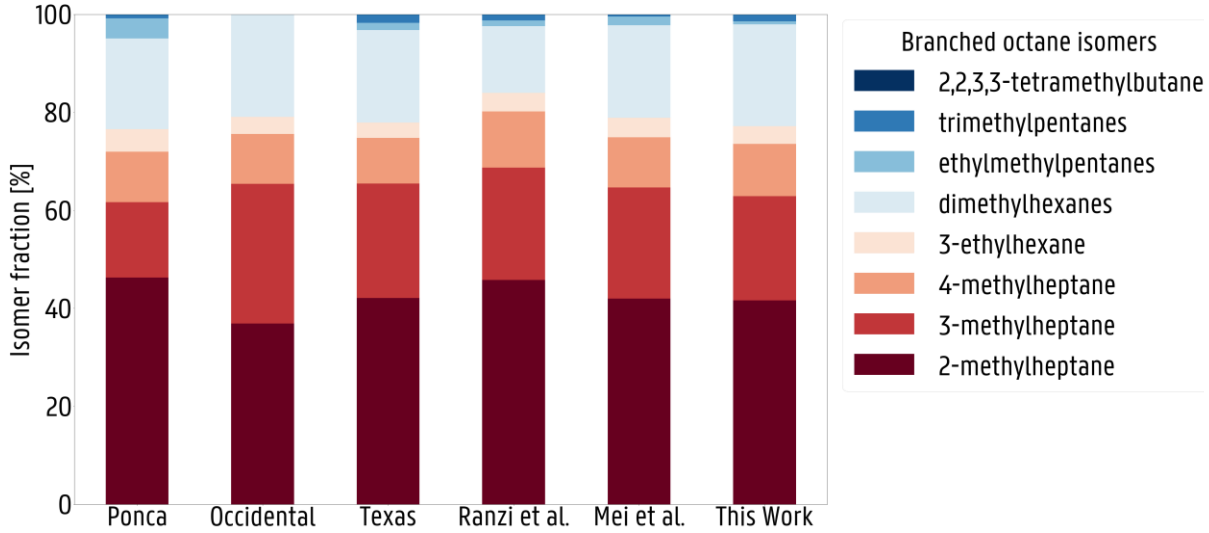


Figure 1: Distribution of branched C_8H_{18} isomers compared to the distributions given by Ranzi *et al.* [56] and Mei *et al.* [57]

Empirical equations, similar to eq **(3)** are constructed for cycloalkanes and aromatic compounds. All rules and factors are given in Supporting Information. The calculation of the

cycloalkane weights is analogous to the calculation for isoparaffins, whereas for aromatic compounds it is taken into account whether the alkyl group is in ortho, meta, or para position. Table 2 gives the molecule-level composition of the Ponca C_4 to C_9 light naphtha fraction experimentally determined [64] and reconstructed with the delumping strategy given above. There is a satisfactory agreement between experimental and predicted values despite the rough approximations of the empirical equations. It is important to notice that the weights and the rules should be revisited, based on experimental evidence when applying the model to renewable feedstocks, since their underlying distributions per lump can differ. Further investigations of heavy and renewable feedstocks with advancements in analytical tools can lead to a better understanding of complex mixtures.

231 **Table 2: Molecule-level composition of the Ponca C4 to C9 light naphtha fraction,**

232 **determined experimentally (Exp) [64] and predicted with delumping rules (Pred)**

IUPAC Name	Exp	Pred	IUPAC Name	Exp	Pred
butane	3.28	3.28	2,6-dimethylheptane	0.18	0.08
pentane	5.83	5.83	2,3-dimethylheptane	0.18	0.16
hexane	6.55	6.55	4-methyloctane	0.36	0.53
heptane	8.37	8.37	2-methyloctane	1.46	0.53
octane	6.92	6.92	3-methyloctane	0.36	0.53
nonane	6.55	6.55	cyclopentane	0.18	0.18
2-methylpropane	1.09	1.09	methylcyclopentane	3.17	3.14
2,2-dimethylpropane	0.00	0.04	cyclohexane	2.58	2.61
2-methylbutane	1.82	1.78	1,1-dimethylcyclopentane	0.58	0.59
2,2-dimethylbutane	0.15	0.06	1,3-dimethylcyclopentane	3.93	3.75
2,3-dimethylbutane	0.29	0.34	1,2-dimethylcyclopentane	1.75	1.88
2-methylpentane	1.35	2.28	ethylcyclopentane	0.58	0.59
3-methylpentane	1.27	0.38	methylcyclohexane	5.83	5.87
2,2-dimethylpentane	0.07	0.05	1,1,3-trimethylcyclopentane	1.09	0.92
2,4-dimethylpentane	0.29	0.31	1,2,4-trimethylcyclopentane	0.84	1.47
2,2,3-trimethylpentane	0.00	0.01	1,2,3-trimethylcyclopentane	1.20	1.47
3,3-dimethylpentane	0.00	0.05	1,1,2-trimethylcyclopentane	0.22	0.18
2,3-dimethylpentane	0.55	0.62	1-methyl-3-ethylcyclopentane	0.44	0.37
2-methylhexane	2.66	2.06	1-methyl-2-ethylcyclopentane	0.66	0.18
3-methylhexane	1.86	2.37	1-methyl-1-ethylcyclopentane	0.11	0.06
3-ethylpentane	0.22	0.17	isopropylcyclopentane	0.04	0.05
2,2,4-trimethylpentane	0.00	0.01	propylcyclopentane	0.22	0.06
2,2,3,3-tetramethylbutane	0.00	0.00	1,4-dimethylcyclohexane	1.24	1.53
2,2-dimethylhexane	0.04	0.04	1,1-dimethylcyclohexane	0.22	0.06
2,5-dimethylhexane	0.22	0.23	1,3-dimethylcyclohexane	2.55	3.06
2,4-dimethylhexane	0.22	0.45	1,2-dimethylcyclohexane	1.35	1.53
2,2,3-trimethylpentane	0.01	0.01	ethylcyclohexane	1.35	0.57
3,3-dimethylhexane	0.11	0.08	benzene	0.55	0.55
2,3,4-trimethylpentane	0.02	0.07	toluene	1.86	1.86
2,3,3-trimethylpentane	0.02	0.01	ethylbenzene	0.69	0.78
2,3-dimethylhexane	0.25	0.45	1,4-dimethylbenzene	0.36	0.58
2-methyl-3-ethylpentane	0.22	0.04	1,3-dimethylbenzene	1.86	1.56
2-methylheptane	3.28	2.95	1,2-dimethylbenzene	0.98	0.97
4-methylheptane	0.73	0.76	isopropylbenzene	0.25	0.20
3,4-dimethylhexane	0.47	0.23	propylbenzene	0.33	0.20
3-methyl-3-ethylpentane	0.07	0.01	1-methyl-3-ethylbenzene	0.62	0.89
3-ethylhexane	0.33	0.25	1-methyl-4-ethylbenzene	0.22	0.55
3-methylheptane	1.09	1.51	1,3,5-trimethylbenzene	0.44	0.57
2,2,4,4-tetramethylpentane	0.00	0.00	1-methyl-2-ethylbenzene	0.33	0.71
2,2,5-trimethylhexane	0.01	0.00	1,2,4-trimethylbenzene	1.86	1.34
2,2,4-trimethylhexane	0.00	0.00	1,2,3-trimethylbenzene	0.69	0.28
2,3,5-trimethylhexane	0.11	0.02			

2.3. Machine Learning Approach

Our machine learning approach consists of four steps that can be summarized as follows. First the generation of the molecular representation is described that serves as input for a neural network. Essential is also the creation of numerical vectors that identify the naphtha mixture. Finally feedforward neural networks are trained to create a regression between input and output vector, and validated using nested cross-validation, a technique to evaluate and optimize the algorithm's performance.

2.3.1. Molecular Vector

Essential is to represent molecules mathematically, as molecules do not have a natural numerical vector representation that can be used as input for artificial neural networks. Therefore, a molecular vector is developed by the Gaussian Learned Histograms of Distances, Angles, and Dihedrals (GauL-HDAD) method [17] for every molecule that is considered in a hydrocarbon mixture. GauL-HDAD is a geometry-based tool, which means that 3D coordinates are needed to set up the molecular representation. In this work, 3D geometries are computed on-the-fly from canonical SMILES identifiers [61, 62] with the ETKDG algorithm in RDKit [44, 69]. The calculated conformer for each molecule is minimized using the Merck Molecular Force Field (MMFF94s) [70]. In the next step, interatomic distances, bond angles, and dihedral angles of all molecules in the dataset are calculated. Using all geometry features of all molecules, histograms are created per individual type of geometry feature, e.g. all carbon-hydrogen interatomic distances are grouped in an CH histogram. Gaussian mixture models (GMM) of all these histograms are created using the unlabeled molecules in the naphtha space, as illustrated in Figure 2. The range of molecules included in the naphtha space determines the application range of the global mixture property prediction tool, in this case *n*-paraffins, iso-

paraffins, olefins, naphthenes and aromatics with up to 12 carbon atoms. The representation of an individual molecule is calculated by first calculating for each geometry feature in a molecule the probability that it is found under any of the gaussians in the GMM. The vectors of all geometry features are then condensed (summed element by element) to a single molecular vector, with the same length as each of the geometry feature vectors, which is the eventual molecular representation. A detailed description of the GauL-HDAD property prediction tool is available in the original paper [17]. The use of the molecular vector representations in the neural network architecture is twofold: (1) for prediction of pure compound properties, and (2) for construction of a condensed mixture representation, which is in turn used to predict mixture properties.

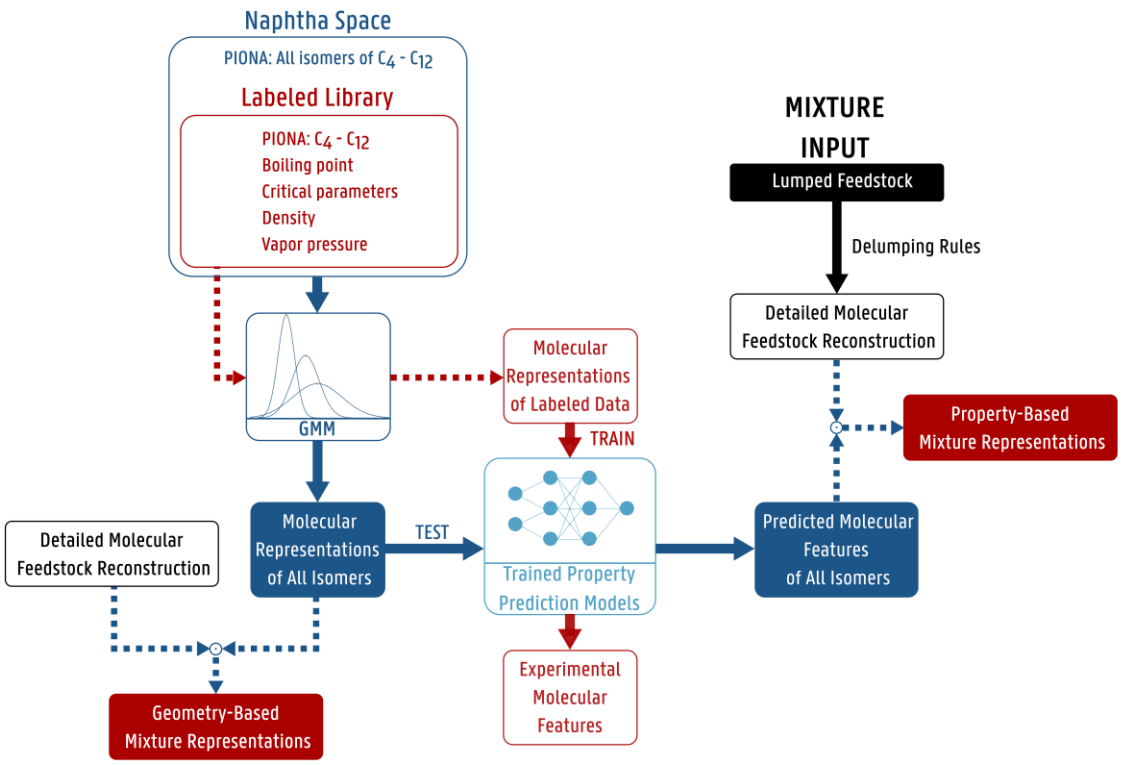


Figure 2: Scheme showing the creation of the two types of mixture representations, starting from a lumped feedstock as identifier of the mixture, which is reconstructed to a molecular composition with delumping rules. Two routes can be distinguished: the blue route contains all (unlabeled) molecules in the naphtha space and is used to set up Gaussian mixture models (GMM). The red route contains the molecules with available property data and is used to train property prediction models. Geometry-based mixture representations are made using the composition and the molecular representations, whereas the property-based mixture representations use the composition and the predicted/experimental molecular features of the individual compounds.

2.3.2. Mixture Representation

Figure 2 illustrates the workflow to create mixture representations. Each sample is represented by two vectors: a geometry-based mixture representation and a property-based mixture representation. The geometry-based mixture representation is a vector that contains information about the constitution of the molecules in the mixture. The earlier introduced molecular

representations are multiplied by the mole fraction of the corresponding molecule in the naphtha sample, as it is obtained from the rule-based molecular reconstruction algorithm. All individual molecular representations are then summed to create one geometry-based mixture representation. The second mixture vector is representative for the properties of the molecules that make up the mixture. If there are no experimental molecular properties available, the properties are predicted using GauL-HDAD from a model trained on the experimental data. Similar to the molecular representations, using reconstructed mole fractions, the molecular features are converted to a single vector with the same length as the feature vector.

279 2.3.3. Feedforward Neural Networks

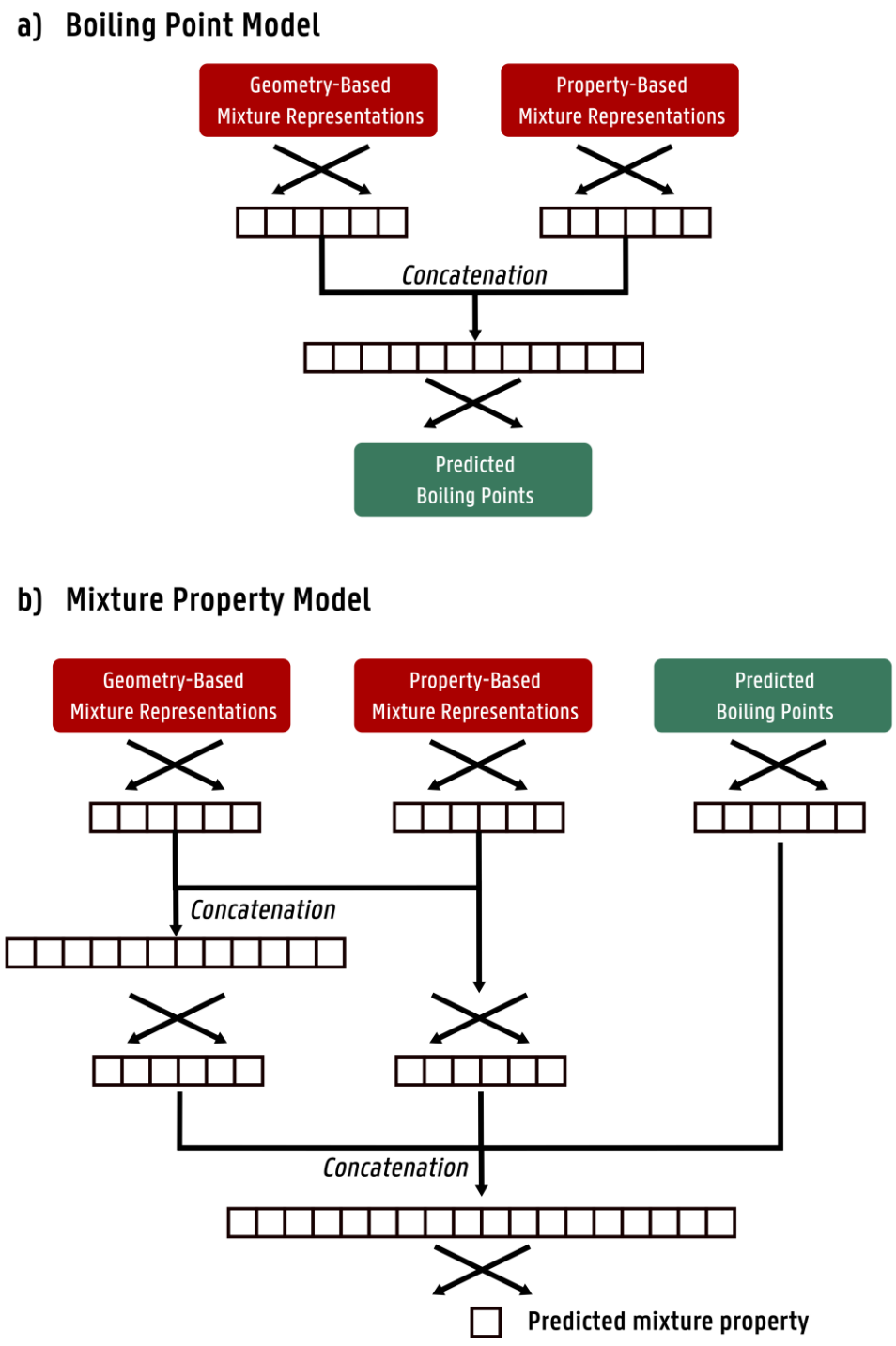


Figure 3: Neural network architecture of the two models. a) The architecture for prediction of boiling points. b) The architecture for mixture properties that depend on the boiling point curve

Figure 3 illustrates the architecture of the feedforward neural network models used in this work for the prediction of mixture properties. The model that predicts points of the distillation curve (Figure 3a, from here on named Boiling Point Model) comprises two input layers and one output layer. The two vectors that represent the mixture – as explained above – are used as the respective input layers. The geometry-based mixture representations and the condensed molecular features are sent through respectively three and two hidden layers, upon which the last hidden layers are concatenated. This concatenated vector is again sent through two hidden layers. The size of the output layer depends on the number of boiling points that are used for training, i.e. three for the data from Pyl *et al.* [49] and eleven for the data from Mei *et al.* [50]. All layers are connected using leaky ReLU activation functions.

The second model (Figure 3b, from here on named Mixture Property Model) predicts properties that are also correlated to the boiling point, such as the density and viscosity. Therefore, this model has a third input layer that contains the predicted boiling points from the Boiling Point Model. The architecture of the mixture property model resembles the boiling point model, but with more complexity due to the additional input. In the mixture property model, hidden layers that are learned versions of the geometry-based mixture representation and of the condensed molecular features are concatenated, and sent through a hidden layer. Additionally, the hidden layer in the condensed molecular feature line goes through a further hidden layer. The third learning line consists of the predicted boiling points, which are sent through hidden layers themselves. Finally, three hidden vectors are concatenated and passed through a hidden layer, yielding as output the desired property. Again, the size of the output depends on the output values chosen for training. In this paper, the output layer size of the mixture property model is always equal to one.

Both models are implemented using Keras [71], integrated in TensorFlow 2 [72]. The neural network parameters are initialized randomly by a normal distribution as published by Glorot et al. [73]. The model was trained using the Adam optimization algorithm, with a fixed learning rate of 0.001 [74]. Training is stopped when 100 epochs are passed in which the validation loss did not decrease. The neural network architectures were selected after a grid search optimization. A complete overview of the architectures of the boiling point and mixture property models is found in Supporting Information and in the source code of the algorithm on GitHub (https://github.com/mrodobbe/naphtha-mixtures).

2.3.4. Nested Cross-Validation

Evaluating all samples is possible using nested cross-validation, also known as double cross-validation. Here, there is an inner and an outer loop. In the outer loop *k* test sets are selected without replacement. In the inner loop *l* validation sets are drawn without replacement for each of the *k* training sets. This means that in total *k* times *l* models are trained for every neural network in this work with *k* different test sets and *k* times *l* different training and validation sets. Each sample in a test set is passed through *l* inner models and the test prediction is the average of the *l* predictions and the uncertainty is the standard deviation of the *l* predictions. The model with the best individual test error in each inner loop is selected as model for the final ensemble of models. In this work, the reported results are for a nested cross-validation algorithm with 10 outer folds and 9 inner folds, which corresponds to an 80/10/10 training/validation/test split. The input and output are shuffled using a seeded random number generator and split into 10 outer folds using the KFold function in scikit-learn [75], so that each datapoint is in exactly 1 outer test set and in 9 outer training sets. This practice is repeated for the inner folds. A datapoint that is in an outer loop training set, will be in 8 inner training sets and in 1 inner validation set.

The implementation of the nested cross-validation algorithm is done using the python package scikit-learn [75].

3. Results and Discussion

3.1. Molecular Property Predictions

The performance of the GauL-HDAD algorithm is tested on six molecular properties, namely the boiling point, critical temperature, density at 293.15 K, vapor pressure at 100 °F (310.93 K), acentric factor, and critical pressure. Parity plots for all properties are given in Figure 4 and an overview of the performance is listed in Table 3. Note that, since the construction of molecular vectors starts from the same Gaussian mixture model, the neural network for the prediction of each of the pure component properties has the same molecular input if the same molecule is considered.

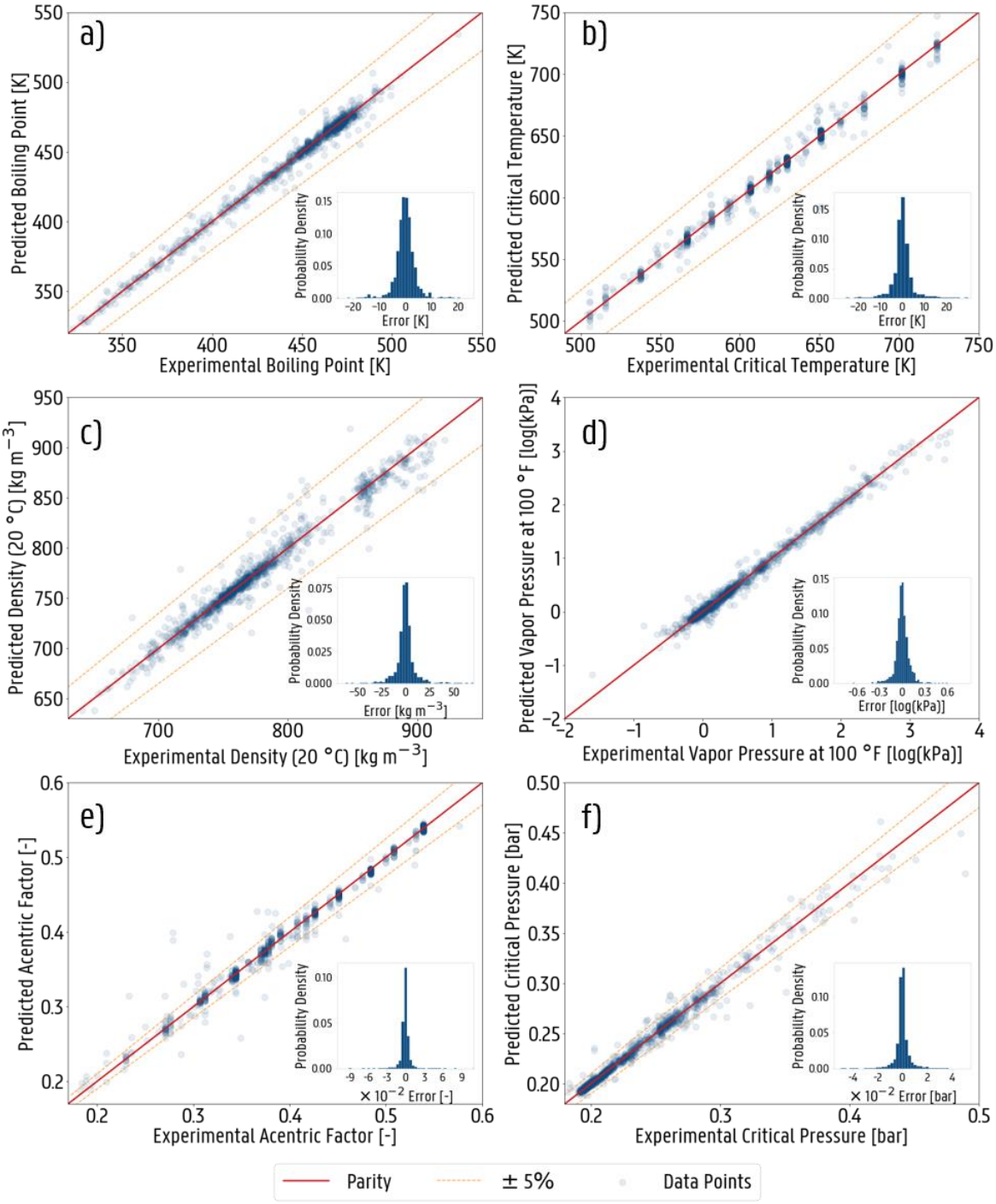


Figure 4: Parity plots for the prediction of the a) molecular normal boiling temperature, b) critical temperature, c) liquid density, d) vapor pressure, e) acentric factor, and f) critical pressure on the test sets of all folds. The orange lines represent the 5% confidence interval.

Among the six pure compound properties, the boiling point data is the most reliable because only experimental datapoints are used for training the model. Densities and vapor pressures are calculated from correlations with experimentally verified parameters. The critical temperature, critical pressure and acentric factor datasets also include calculated values, because experimental values are too scarce (less than 100 datapoints) for these properties. The effect of calculated data is mainly visible in the parity plots of the critical temperature (Figure 4b) and the acentric factor (Figure 4e), by the vertical lines. These vertical lines are several isomers which have the same listed, calculated value but for which the machine learning algorithm predicts different values. This illustrates the difference between group contribution methods and machine learning methods. The group contribution method uses empirical values per functional group, but it fails at distinguishing structural isomers.

Table 3: Performance of different pure compound prediction models

Property	**Unit**	**MAE**	**RMSE**	**R^2**
Normal boiling point	*K*	2.5	3.9	0.992
Critical temperature	*K*	2.8	5.2	0.990
Liquid density	$kg\ m^{-3}$	5.9	10.2	0.965
Vapor pressure	*log(kPa)*	0.060	0.092	0.988
Acentric factor	-	0.0046	0.010	0.988
Critical pressure	*bar*	0.34	0.69	0.979

The boiling point predictions (Figure 4a) have the highest R^2 value of all properties and only four molecules have an error higher than 5%. These four molecules are all unsaturated molecules. Three of these molecules contain a cyclohexane ring with an unsaturated group and the other molecule is a cumulated diene. Unsaturated naphthenes and cumulated dienes are predicted with large uncertainties for all properties, which indicates data scarcity of these molecular classes in the dataset. However, the impact in the rest of the algorithm is limited since unsaturated naphthenes and cumulated dienes are excluded using the current molecular

reconstruction scheme, because of their negligible occurrence in naphtha samples. It is clear that accurate data for these olefinic compounds would be extremely useful because for example in plastic waste derived naphthas the olefin content is substantially larger than for fossil naphthas. Given the fact that correlation between boiling points and critical temperatures have been proposed [1], it is not surprising that the performance on predicting critical temperatures is similar than for boiling points. This is also true for Pitzer's acentric factor (Figure 4e), which uses by definition the critical temperature, critical pressure and vapor pressure [76]. The liquid density of pure compounds (Figure 4c) is predicted less accurate than the other properties. A comparison to experimental data learns that in particular predictions of cyclic molecules are poorer than for acyclic molecules. The reduced agreement with experimental data can be linked to the higher density values of these cyclic compounds compared to acyclic molecules. In Figure 4c, this is visible by the overpredicted values with an experimental value around 800 kg m^{-3} and the underpredicted values at the higher end of the range.

Table 4 shows a comparison between molecular property predictions with the machine learning tool GauL-HDAD and the Joback-Reid group contribution method (GC) [6]. The GC values are calculated using the python package JRGui [77]. The machine learning predictions are test set values from the models reported above. A total of 86 samples is taken, for which experimental normal boiling points, critical temperatures, and critical pressures are available. The samples include 9 n-paraffins, 36 isoparaffins, 11 olefins, 14 naphthenes, and 16 aromatics. It is observed that the machine learning method outperforms the GC method for all three physical properties. Only for the normal boiling temperature the performance is similar to what is reported in Table 3. This behavior is due to the fact that the normal boiling point model is trained only on experimental data, while the critical temperature and pressure models are trained

also on GC data. Only experimental data is used in the comparison in Table 4. This explains why the performance of these models is closer to the GC performance.

Table 4: Comparison of machine learning predictions with GauL-HDAD and Joback-Reid group contribution (GC) calculations with experimental normal boiling points, critical temperatures, and critical pressures of 86 molecules.

Property	Model	Unit	MAE	RMSE	R^2
Normal boiling point	*GauL-HDAD*	*K*	2.7	4.3	0.993
Normal boiling point	*GC*	*K*	9.5	11.8	0.952
Critical temperature	*GauL-HDAD*	*K*	7.9	10.5	0.976
Critical temperature	*GC*	*K*	15.9	18.97	0.915
Critical pressure	*GauL-HDAD*	*bar*	1.58	2.48	0.829
Critical pressure	*GC*	*bar*	1.89	2.81	0.809

3.2. Mixture Property Prediction

The property-based mixture representation consists of the normal boiling point, the critical temperature, the liquid density, and the vapor pressure. If experimental data is available for a molecule, then the experimental datapoints are used. For most of the compounds in a naphtha, experimental data is not available and is predicted using the above discussed machine learning tool. The property-based mixture representation is composed of a fifth value: the carbon-to-hydrogen ratio, which is directly calculatable from the molecular structure. These property-based mixture representations are made by multiplying the property vector per molecule with the absolute fraction of that molecule, followed by summing the vectors to a single vector. Two models are trained in sequence: the boiling point model links the geometry-based and property-based mixture representations to a boiling point curve, and the mixture property model links the geometry-based and property-based mixture representations, and the boiling point curve to bulk properties. An overview of the performance of the two models is provided in Table 5.

Table 5: Performance of the boiling point model (BPM) and the mixture property prediction model (MPM)

Property	Ref. Data	Unit	Model	MAE	RMSE	R^2
Boiling point curve	[49]	*K*	BPM	2.4	3.8	0.996
	[50]	*K*	BPM	2.9	4.0	0.991
Specific gravity	[49]	-	MPM	8.2×10^{-4}	1.1×10^{-3}	0.988
Liquid density	[50]	$kg\ m^{-3}$	MPM	2.3	2.6	0.958
Dynamic viscosity	[50]	*Pa.s*	MPM	7.2×10^{-3}	8.2×10^{-3}	0.975
Surface tension	[50]	$N\ m^{-1}$	MPM	1.2×10^{-4}	1.5×10^{-4}	0.963

3.2.1. Boiling Point Curve Prediction

The largest dataset of naphthas contains 272 samples with an initial boiling point (IBP), a boiling point at 50% (BP50), and a final boiling point (FBP) experimentally measured according to the ASTM D86 method [36] and converted to true boiling points using Riazi's correlations [51]. Figure 5 shows the parity plot for the boiling point model on this dataset. Across all the boiling points, an MAE of 2.4 K and an RMSE of 3.8 K is achieved, which is about similar to the error on the pure compound boiling points. As can be seen in Figure 5, the errors should be evaluated individually. The performance is best for the BP50 prediction, with an MAE of 1.5 K and an RMSE of 2.15 K. Predictions of the IBP are in the same order with an MAE of 1.8 K and an RMSE of 2.6 K, but it should be remarked that the data range of the IBP is much smaller than that of the BP50. The FBP is predicted with the largest errors, namely an MAE of 3.9 K and an RMSE of 5.6 K. There are several considerations that should be made along with the IBP and FBP predictions. First of all, it is hard to experimentally measure the IBP and the FBP because the naphtha samples might contain very volatile compounds with a boiling point lower than ambient temperature (e.g. propane) or heavy products with low volatility that do not evaporate. In that sense, it is hard to select the initial and final boiling points. A common way to avoid the problem is by taking the boiling point at 2 or 5% and 95 or

98% as respective IBP and FBP [51]. A second noise on the predictions is the actual composition. Even though a complete reconstruction is made, the initial lumped composition contains experimental noise by itself. This is because lumps need to be estimated from GCxGC chromatograms with a mass balance that is usually not fully closed. The errors on all boiling points are lower than predicted by Plehiers *et al.* [37], who achieved RMSEs of 3.5, 2.7, and 5.7 K on respectively the IBP, BP50, and FBP. However, the comparison is not on the same set since Plehiers *et al.* did not use cross-validation in their work.

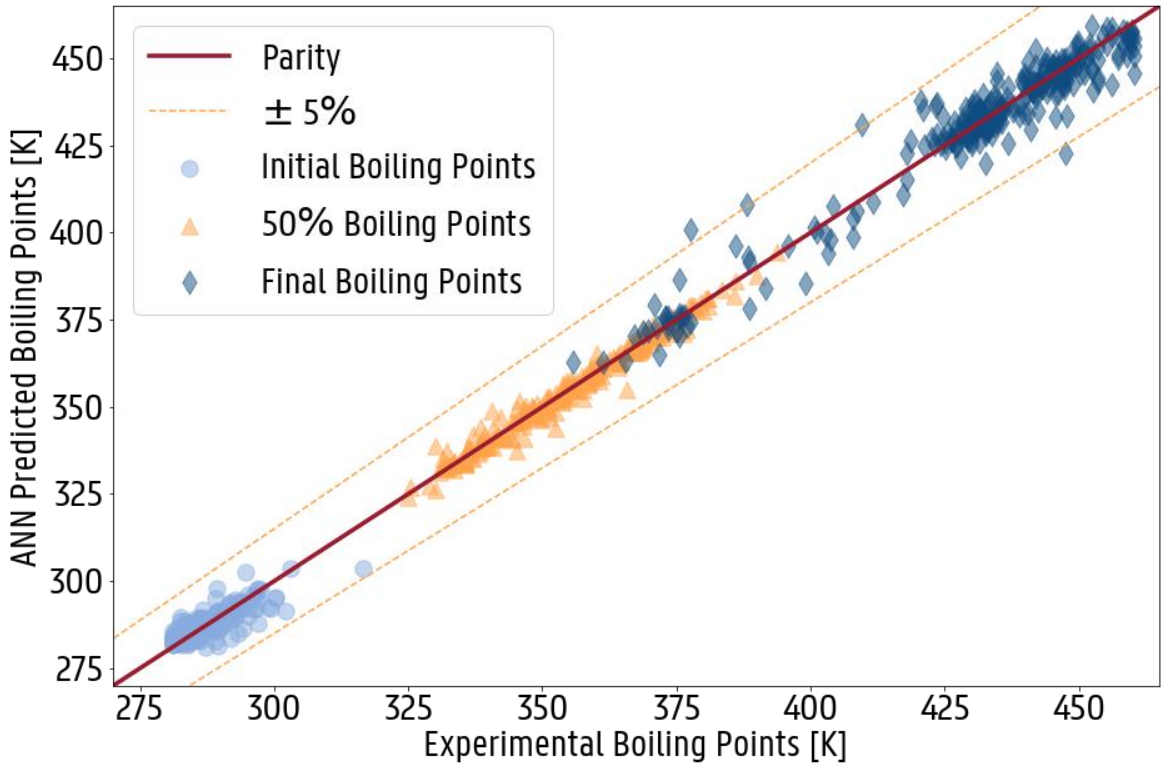


Figure 5: Performance of the boiling point model on IBP, BP50 and FBP predictions of 272 naphthas using nested cross-validation on the test sets of all folds.

Similar performance is observed for the other, smaller dataset with 110 naphtha samples. The average performance is slightly worse with an MAE of 2.9 K and an RMSE of 4.0 K, and the parity plot is shown in Figure 6. In this dataset also the boiling point at 10%, 20%, 30%, 40%,

50%, 60%, 70%, 80%, 90%, 95% have been measured. The errors are, in agreement with the previous results, larger for BP10 and BP95 with respectively MAEs of 3.5 K and 5.5 K, and RMSEs of 4.4 K and 7.6 K. The other boiling points have relatively similar errors to each other, and are all much lower than the 5% error range which corresponds to ±20 K at 400 K. The MAEs of the intermediary boiling points range from 2.0 to 3.1 K and the RMSEs from 2.4 to 4.2 K.

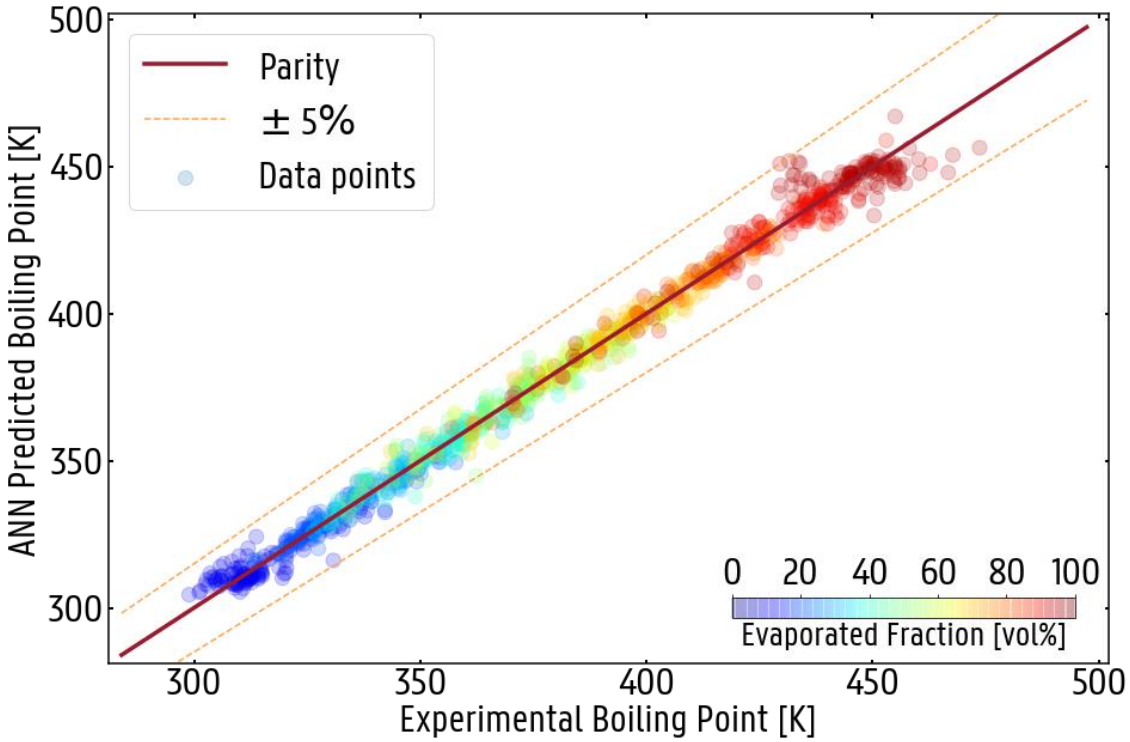


Figure 6: Parity plot with all test set boiling points in the Mei dataset, predicted with the boiling point model. The color of the point indicates to which part of the distillation curve the boiling point belongs.

The errors can be divided into two classes: samples with high model uncertainty and high prediction errors, samples with low model uncertainty and high prediction errors. The model is quite uncertain (i.e. the individual ensemble models disagree resulting in a high variance on the predictions), when the input is quite distinctive from the rest of the training set. This indicates that the experimental value is likely correct, but that the model is used outside its application

range. In the second class of errors, the model has a rather low model uncertainty and the sample resembles the training samples. In this case, the model is used within its application range and we can assume that there is a significant experimental error on the sample. The three samples with the highest prediction variance for specific gravity predictions in the Pyl dataset, are analyzed by assessing the degree of similarity with the other naphthas. A principal component analysis is performed on the geometry-based mixture representations of 269 naphthas from the Pyl dataset [49]. The first three principal components, which explain more than 90% of the variance, are shown in Figure 7. The ellipsoid in Figure 7 corresponds to a Mahalanobis distance of about 2.5, which means that it encloses 90% of the data. The Mahalanobis distance has been used in previous work to show similarities between naphtha samples [37, 65]. It is seen from Figure 7 that the three samples with the highest variance (red spheres) have a Mahalanobis distance larger than the critical distance of 2.5. Nevertheless, these samples are not the only samples that are out of the application range, but the machine learning approach manages to achieve highly accurate predictions.

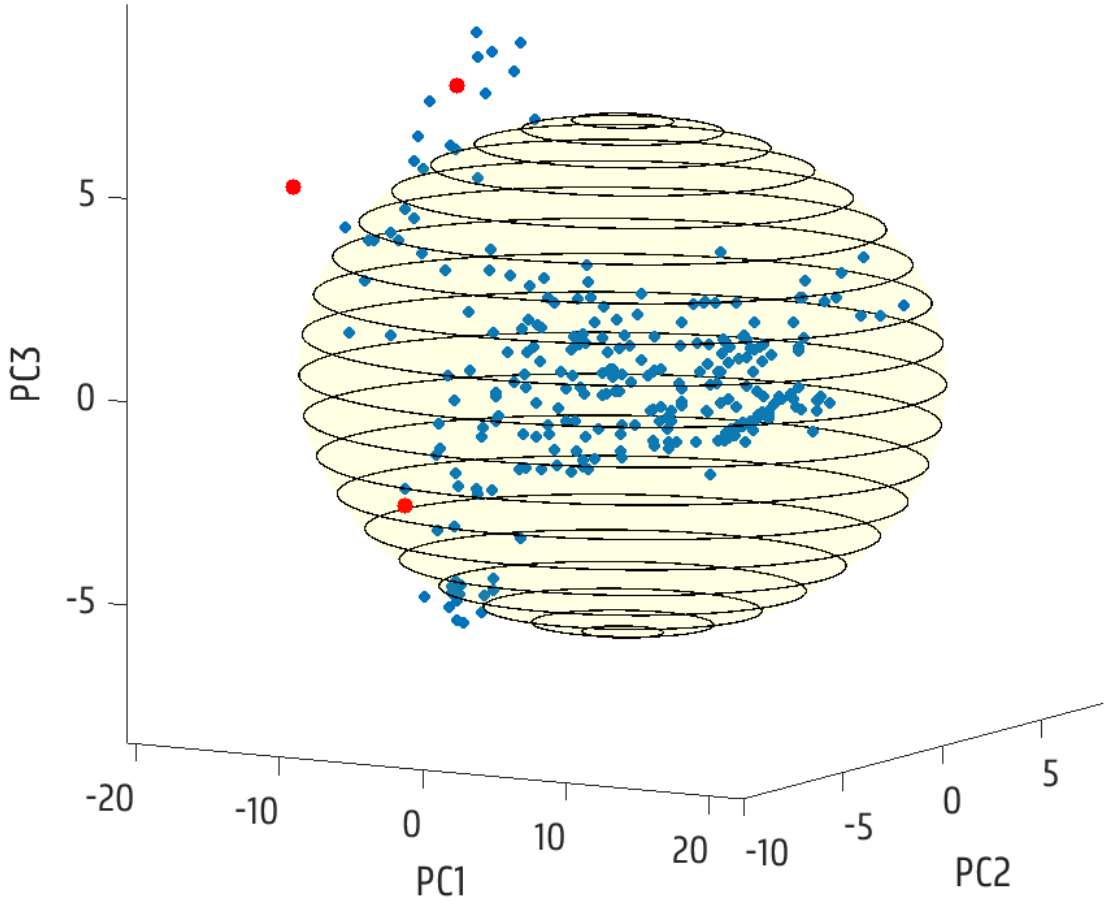


Figure 7: First three principal components (PC1, PC2, PC3) of the geometry-based mixture representation of the naphthas in the Pyl database. The red spheres are the three naphthas for which the variances is the largest on specific gravity predictions.

452

453

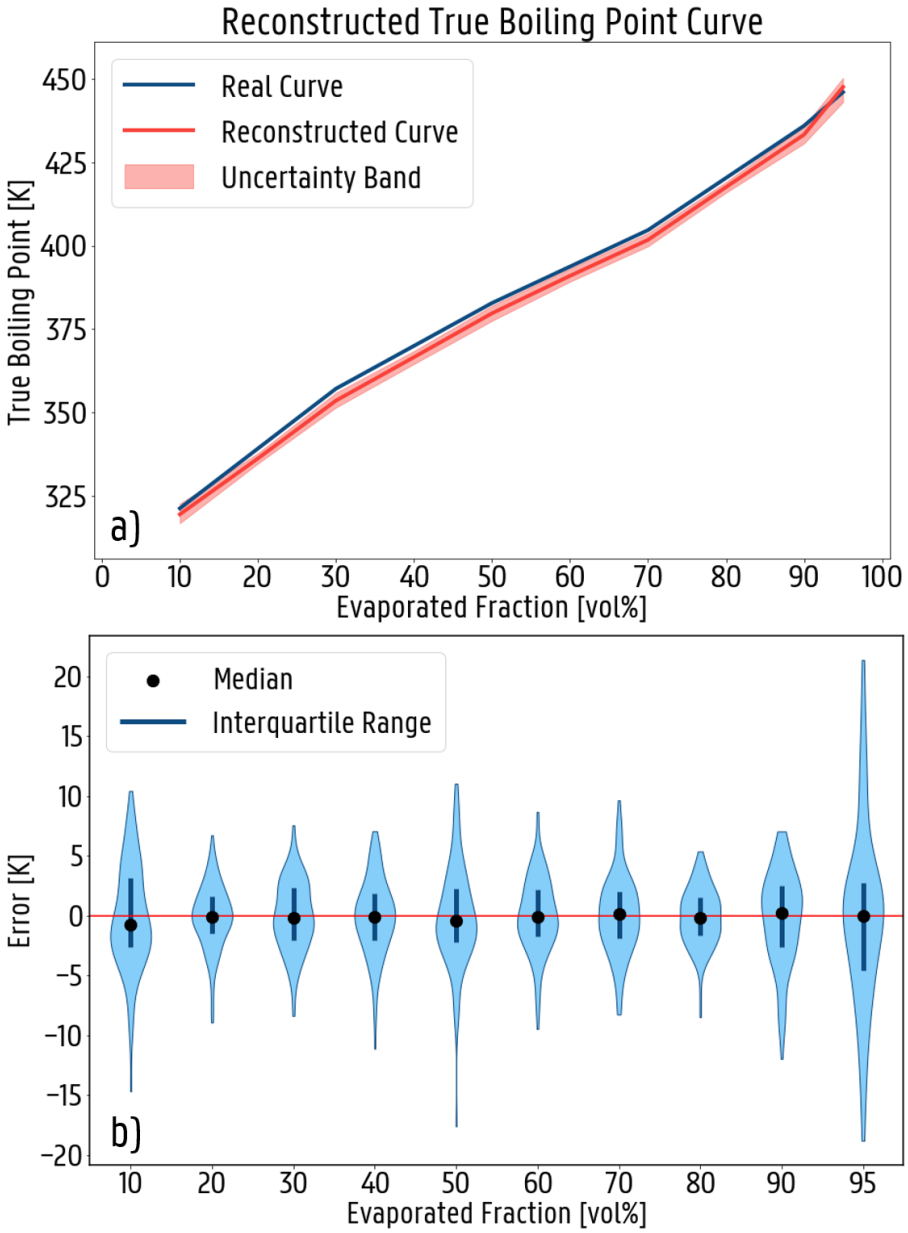


Figure 8: a) The boiling point curve of a representative naphtha sample predicted by the boiling point model, compared to the true curve. The shaded area indicates the uncertainty on the predictions of each point. b) Error distribution of the mixture boiling point predictions in the Mei dataset by the boiling point model as function of the evaporated volume with the median error and the interquartile range indicated.

Figure 8a goes in more detail than Figure 6 by showing the individual boiling point curve of a naphtha sample. The naphtha sample (sample 67 from the Mei dataset) is chosen to be representative, with its MAE close to the median value of all MAEs. The trend is clearly

457 captured in the prediction of the boiling point curves and shows that a complete boiling point
458 curve can be reconstructed with this approach. Figure 8b confirms the larger errors on the 10%
459 and 95% boiling points, which were also noticed in Figure 6. Errors at the lower and higher end
460 of the distillation curve can be related to the experimental difficulty to measure the boiling point
461 at these fractions. The median error is in all cases very close to zero and the middle 50% of the
462 errors is found within the range from -5 to 5 K.

463 ### 3.2.2. Prediction of Specific Gravity, Liquid Density, Dynamic Viscosity, and Surface
464 Tension

465 The mixture property model predicts bulk properties from three input vectors and different
466 versions are trained on specific gravities, liquid densities, dynamic viscosities, and surface
467 tensions. The high R^2 values listed in Table 5 show a very good agreement between the true
468 and the predicted values.

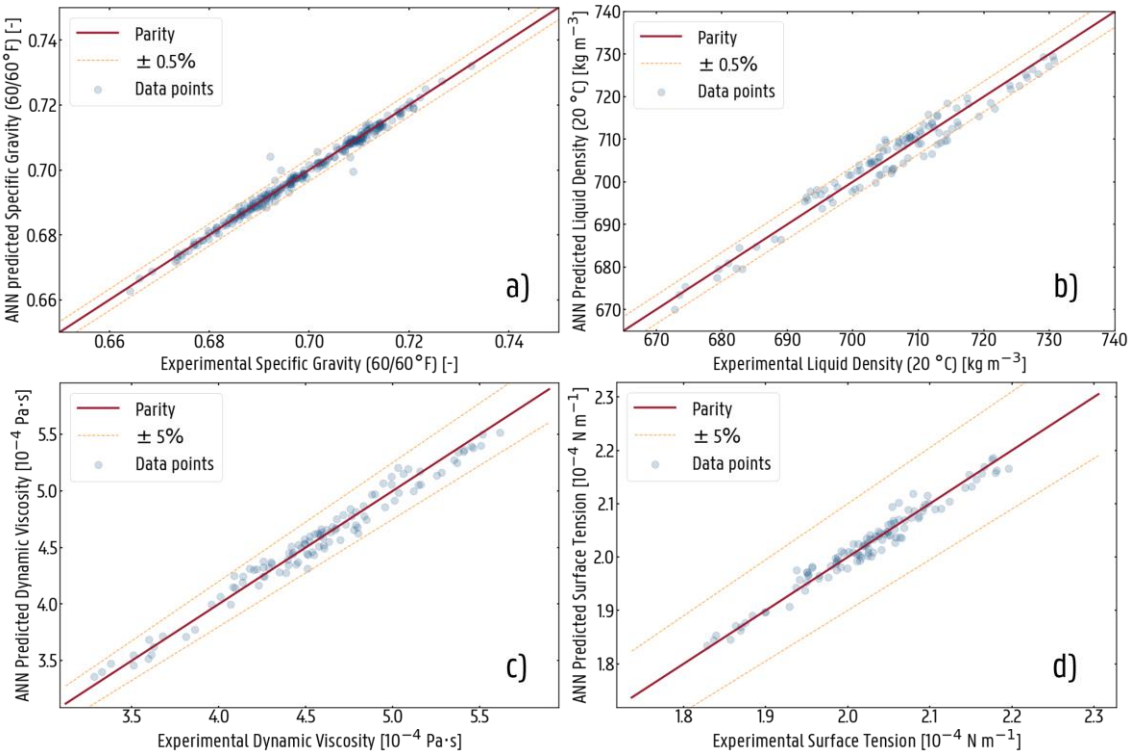


Figure 9: Performance of the mixture property model on predicting a) the specific gravity in the Pyl database, b) the liquid density in the Mei database, c) dynamic viscosity in the Mei database, d) the surface tension in the Mei database on the test sets of all folds. The orange line indicates the 0.5% confidence interval.

Figure 9a and Figure 9b show the parity plots for the mixture property model predictions of specific gravity and liquid density. For both datasets, the prediction is highly accurate with most data points found within 0.5% error. The values with larger errors are samples of which the boiling point predictions have larger errors too, which is likely an effect of an incorrect lumped composition.

Figure 9c and Figure 9d show the parity plots for the prediction of the dynamic viscosity and the surface tension of the naphtha samples in the dataset of Mei *et al.* [50]. Both properties are estimated with an excellent accuracy. This result is remarkable because the property-based mixture representation does not contain any viscosity nor surface tension values of individual components. This means that it is possible learn and predict physicochemical properties of

479 molecular mixtures without the need for experimental values of that property for the individual
480 pure components.

481 ### 3.2.3. Comparison with Kay's Mixing Rule

482 The common approach in molecular reconstruction tools is using linear mixing rules, also
483 known as Kay's mixing rules, given by eq (5) [51].

$$\theta_m = \sum_i x_i \theta_i \quad \textbf{(5)}$$

484 These correlations are assumed to obtain high accuracy when the composition is known and
485 when the individual properties are known. In the molecular reconstruction of naphthas,
486 obtaining an accurate composition is not straightforward and neither is having accurate
487 properties. Especially when considering alternative feedstocks, such as plastic waste pyrolysis
488 oils, it is currently not possible to have a composition in high detail [40]. In addition,
489 experimental property values of species involved in those feedstocks, such as branched olefins,
490 are mostly not available.
491 Figure 10 shows the parity plot when calculating the liquid density using Kay's mixing rule.
492 The fractions in eq (5) are the same as used in the boiling point model and the mixture property
493 model, and the values are predicted using GauL-HDAD. It is clearly visible that the
494 performance of the mixing rule is very much equal to the performance by the mixture property
495 model, shown in Figure 9b. The RMSE of the linear model is 2.9 kg m^{-3} compared to 2.6 kg m^{-3}
496 of the mixture property model. The R^2-value is 0.952, which is almost equal to the 0.956 of the
497 mixture property model. Rather than one trend in Figure 9b, two parallel trends are noticed in
498 Figure 10. The systematically overestimated naphtha samples are richer in aromatics and
499 naphthenes, whereas the underestimated data are richer in (iso)paraffines. It is not surprising
500 that Kay's mixing rule for density is accurate. Experimental density values are present for the

large majority of the molecules in the naphtha samples and, as shown in Figure 4, the density predictions are reliable with training values available over the whole range of data points. Moreover, the performance of the linear mixing rule is a proof that the molecular reconstruction rules presented in this paper are reliable for naphthas.

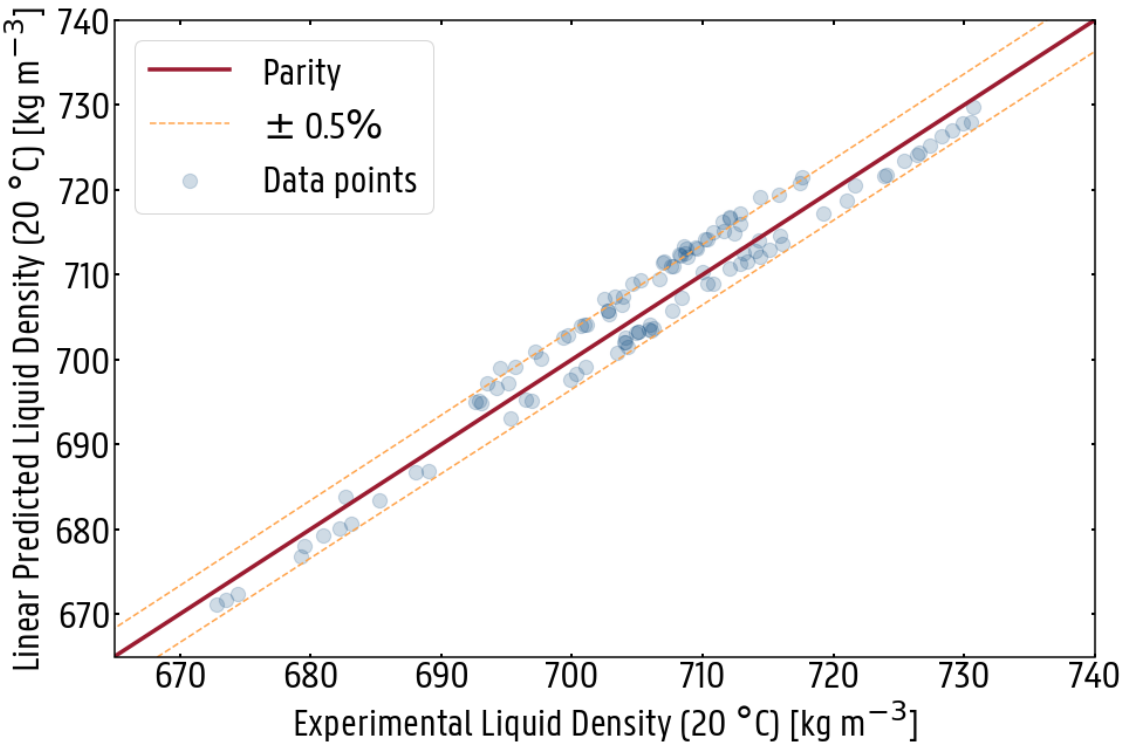


Figure 10: Parity plot of the liquid density in the Mei dataset using the linear mixing rules

It is more interesting to consider a mixture property for which less data is available, such as the dynamic viscosity at 298.15 K. The dynamic viscosity is a transport property, but also a property that is a crucial quality parameter, such as for lubricants. Because it is so temperature-dependent, it is very hard to quantify experimentally which makes accurate computational models of high importance. GauL-HDAD is trained on 389 experimental and calculated dynamic viscosity values at 298.15 K for pure components from Yaws' Handbook of Transport Properties [78]. Similar to the pure compound properties above, the considered molecules are all from the naphtha space. The first consideration to be made is that the viscosity values are

not evenly distributed in the training set. Even though the number of isomers grows with the number of carbon atoms, the number of molecules with experimental viscosity values decreases above 10 carbon atoms. Only a handful of values are available for molecules with 12 carbons. The sparse distribution of data in the training set is reflected by the performance of GauL-HDAD. A parity plot is shown in Figure 11a. The largest outliers are the C_9 to C_{12} monoalkylcyclohexanes, which are also outliers in the distribution (the largest value per carbon number). Since these monoalkylcyclohexanes are nearly the only naphthenics in the database with higher carbon numbers and because naphthenics are important in naphtha, this makes the predictions unreliable. Yet, GauL-HDAD performs very accurately disregarding these outlier values.

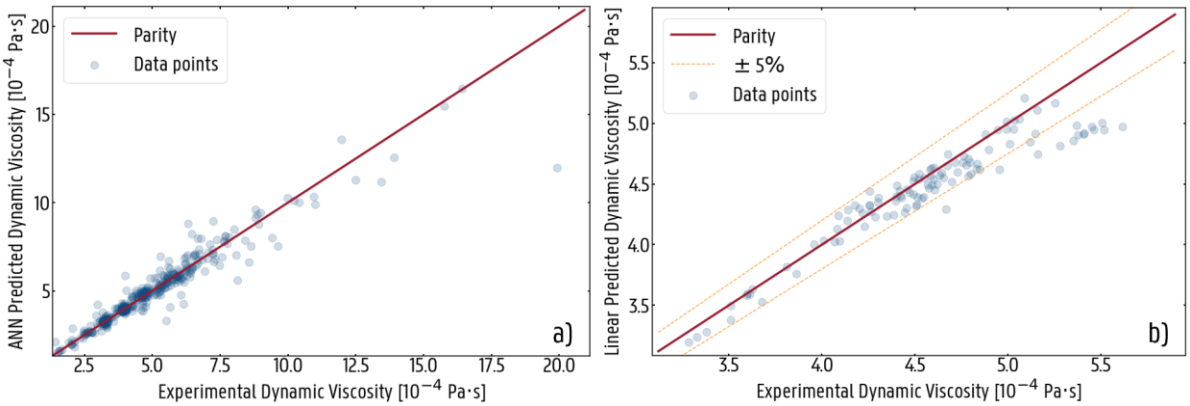


Figure 11: a) Performance of pure compound dynamic viscosity values using GauL-HDAD. b) Parity plot of the dynamic viscosity in the Mei dataset using linear mixing rules

The effect of unreliable pure compound properties is clearly visible in Figure 11b. Using The mixture property model, the R^2 value was 0.975, which dropped to 0.899 for Kay's mixing rules. Although still acceptable, an underestimation of the higher viscosity values is witnessed. Because the density can be predicted properly with the same absolute fractions, the cause of the poorer predictions is the unreliable prediction of the individual viscosity values. Especially for

heavier mixtures, e.g. for lubricants, the viscosity is important. Knowing that viscosity increases with molecular weight, this shows the large benefit of the neural network-based model.

The main advantage of the algorithm that integrates pure compound prediction, boiling point curve prediction, and bulk property prediction lies in the fact that bulk properties can be predicted without the need for experimental values of that property. In this way, there are similarities with equations of state that are used in process simulation software today. The new approach suits well in these process simulators because it can predict properties from the composition, whereas predictions are now made in the other way to yield a detailed composition. The software can also applied for feeds with higher carbon numbers, when data is available, since two-dimensional gas chromatography is a powerful tool for analyzing heavy feedstocks [39, 65, 79-81]. When considering heavier feedstocks, it is needed to represent lumps by model compounds since the number of isomers becomes astronomically large and because current analytical techniques cannot characterize heavy feeds to such a level of detail [82]. When data is provided of feeds with higher carbon numbers, such as diesel or vacuum gasoil, the neural networks can be retrained to extend the application range. The main limitation is the lack of large datasets in this field. Large amounts of data are available in academia and industry, which have not yet been made public. This data-driven tool can speed up the optimization of processes with new feedstocks when new datasets are made available. Apart from the energy industry, it is generally important to have accurate predictions of molecular mixtures. This work shows that it is possible to link an averaged molecular structure to a mixture property. Future work should evaluate property prediction of mixtures with known compositions as well as of renewable feedstocks.

4. Conclusions

In this work, we develop a machine learning algorithm to predict boiling points and bulk properties of naphthas, starting from a lumped composition. It is found that a delumped molecular composition of the mixture components and general physical properties from pure components, such as the liquid density and the critical parameters, are sufficient to predict properties of complex hydrocarbon mixtures. Linear mixing rules, which are typically applied in industry, can only perform well when accurate property estimates of the pure compounds are available. Therefore the neural network-based property prediction tool GauL-HDAD was trained on normal boiling points, critical temperatures, critical pressures, liquid densities, acentric factors, vapor pressures, and viscosities of pure hydrocarbons. The predicted values of densities and viscosities for pure components are applied in linear mixing rules to predict mixture properties, and it is seen that only mixture densities are predicted accurately because only a very small amount of experimental viscosity values for pure components are available. We have developed a neural network-based approach that can successfully predict boiling point curves, densities, viscosities, and surface tensions of mixtures. The common factor in the property prediction of pure compound properties and bulk properties is a molecular representation vector that captures the inner structure of the molecule, so that the mixture can be regarded as a pseudo-molecule. The neural network-based approach has a second input vector that is made from pure component predictions. Scientific progress in characterization of renewable feedstocks and of heavier mixtures can provide a better understanding of the composition of these mixtures and more experimental data. With the availability of new data, the newly developed algorithms can become a reliable tool to predict mixture properties of

naphthas with slightly different compositions and, hence, speed up the design of new chemical processes.

Supporting Information:

S1: Code available as free software under the MIT license on https://github.com/mrodobbe/naphtha-mixtures. S2: Delumping rules per class. S3: ANN architectures. S4: Pure compound properties. S5: Description of principal component analysis. This information is available free of charge via the Internet at https://pubs.acs.org/.

Acknowledgement:

Maarten Dobbelaere and Yannick Ureel acknowledge financial support from the Fund for Scientific Research Flanders (FWO Flanders) respectively through doctoral fellowship grants 1S45522N and 1185822N. Florence Vermeire acknowledges a personal grant from the Research Fund of Ghent University (BOF). The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme / ERC grant agreement n° 818607. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI

5. References

1. Poling, B.E.; Prausnitz, J.M.; O'connell, J.P. *Properties of gases and liquids*. McGraw-Hill Education, **2001**.
2. Ren, Y.; Liao, Z.; Sun, J.; Jiang, B.; Wang, J.; Yang, Y.; Wu, Q. Molecular reconstruction: Recent progress toward composition modeling of petroleum fractions. *Chemical Engineering Journal* **2019,** *357*, 761-775.
3. Deniz, C.U.; Yasar, S.H.O.; Yasar, M.; Klein, M.T. Effect of Boiling Point and Density Prediction Methods on Stochastic Reconstruction. *Energy & Fuels* **2018,** *32*, 3344-3355.
4. Peschel, A.; Freund, H.; Sundmacher, K. Methodology for the Design of Optimal Chemical Reactors Based on the Concept of Elementary Process Functions. *Industrial & Engineering Chemistry Research* **2010,** *49*, 10535-10548.
5. Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, M.; Kahn, I.; Dobchev, D.A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews* **2010,** *110*, 5714-5789.
6. Joback, K.G.; Reid, R.C. Estimation of Pure-Component Properties from Group-Contributions. *Chemical Engineering Communications* **1987,** *57*, 233-243.
7. Hukkerikar, A.S.; Sarup, B.; Ten Kate, A.; Abildskov, J.; Sin, G.; Gani, R. Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilibria* **2012,** *321*, 25-43.
8. Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria* **2001,** *183-184*, 183-208.
9. Constantinou, L.; Gani, R.; O'Connell, J.P. Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. *Fluid Phase Equilibria* **1995,** *103*, 11-22.
10. Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE Journal* **1994,** *40*, 1697-1710.
11. Stein, S.E.; Brown, R.L. Estimation of normal boiling points from group contributions. *Journal of Chemical Information and Computer Sciences* **1994,** *34*, 581-587.
12. Dobbelaere, M.R.; Plehiers, P.P.; Van de Vijver, R.; Stevens, C.V.; Van Geem, K.M. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. *Engineering* **2021,** *7*, 1201-1211.
13. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O.A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012,** *108*, 058301.
14. Faber, F.A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S.S.; Dahl, G.E.; Vinyals, O.; Kearnes, S.; Riley, P.F.; von Lilienfeld, O.A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **2017,** *13*, 5255-5264.
15. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. *et al.* Analyzing Learned Molecular Representations

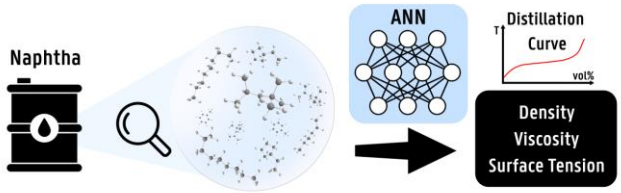
for Property Prediction. *Journal of Chemical Information and Modeling* **2019,** *59*, 3370-3388.
16. Yalamanchi, K.K.; van Oudenhoven, V.C.O.; Tutino, F.; Monge-Palacios, M.; Alshehri, A.; Gao, X.; Sarathy, S.M. Machine Learning To Predict Standard Enthalpy of Formation of Hydrocarbons. *The Journal of Physical Chemistry A* **2019,** *123*, 8305-8313.
17. Dobbelaere, M.R.; Plehiers, P.P.; Van de Vijver, R.; Stevens, C.V.; Van Geem, K.M. Learning Molecular Representations for Thermochemistry Prediction of Cyclic Hydrocarbons and Oxygenates. *The Journal of Physical Chemistry A* **2021,** *125*, 5166-5179.
18. Grambow, C.A.; Li, Y.-P.; Green, W.H. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *The Journal of Physical Chemistry A* **2019,** *123*, 5826-5835.
19. Vermeire, F.H.; Green, W.H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021,** *418*, 129307.
20. Chung, Y.; Vermeire, F.H.; Wu, H.; Walker, P.J.; Abraham, M.H.; Green, W.H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *Journal of Chemical Information and Modeling* **2022,** *62*, 433-446.
21. Schweidtmann, A.M.; Rittig, J.G.; König, A.; Grohe, M.; Mitsos, A.; Dahmen, M. Graph Neural Networks for Prediction of Fuel Ignition Quality. *Energy & Fuels* **2020,** *34*, 11395-11407.
22. vom Lehn, F.; Brosius, B.; Broda, R.; Cai, L.; Pitsch, H. Using machine learning with target-specific feature sets for structure-property relationship modeling of octane numbers and octane sensitivity. *Fuel* **2020,** *281*, 118772.
23. Chalk, A.J.; Beck, B.; Clark, T. A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *Journal of Chemical Information and Computer Sciences* **2001,** *41*, 457-462.
24. Oprisiu, I.; Marcou, G.; Horvath, D.; Brunel, D.B.; Rivollet, F.; Varnek, A. Publicly available models to predict normal boiling point of organic compounds. *Thermochimica Acta* **2013,** *553*, 60-67.
25. Zang, Q.; Mansouri, K.; Williams, A.J.; Judson, R.S.; Allen, D.G.; Casey, W.M.; Kleinstreuer, N.C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *Journal of Chemical Information and Modeling* **2017,** *57*, 36-49.
26. Santak, P.; Conduit, G. Predicting physical properties of alkanes with neural networks. *Fluid Phase Equilibria* **2019,** *501*, 112259.
27. Wessel, M.D.; Jurs, P.C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *Journal of Chemical Information and Computer Sciences* **1995,** *35*, 68-76.
28. Alshehri, A.S.; Tula, A.K.; You, F.; Gani, R. Next generation pure component property estimation models: With and without machine learning techniques. *AIChE Journal* **2021,** *n/a*, e17469.
29. Stratiev, D.; Shishkova, I.; Tankov, I.; Pavlova, A. Challenges in characterization of residual oils. A review. *Journal of Petroleum Science and Engineering* **2019,** *178*, 227-250.
30. Grunberg, L.; Nissan, A.H. Mixture Law for Viscosity. *Nature* **1949,** *164*, 799-800.

31. Monnery, W.D.; Svrcek, W.Y.; Mehrotra, A.K. Viscosity: A critical review of practical predictive and correlative methods. *The Canadian Journal of Chemical Engineering* **1995,** *73*, 3-40.
32. Roelands, C.J.A.; Vlugter, J.C.; Waterman, H.I. The viscosity-temperature-pressure relationship of lubricating oils and its correlation with chemical constitution. *Journal of Basic Engineering* **1963,** *85*, 601-607.
33. Lohrenz, J.; Bray, B.G.; Clark, C.R. Calculating Viscosities of Reservoir Fluids From Their Compositions. *Journal of Petroleum Technology* **1964,** *16*, 1171-1176.
34. Abdul Jameel, A.G.; Van Oudenhoven, V.; Emwas, A.-H.; Sarathy, S.M. Predicting Octane Number Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks. *Energy & Fuels* **2018,** *32*, 6309-6329.
35. Albahri, T.A. Specific Gravity, RVP, Octane Number, and Saturates, Olefins, and Aromatics Fractional Composition of Gasoline and Petroleum Fractions by Neural Network Algorithms. *Petroleum Science and Technology* **2014,** *32*, 1219-1226.
36. ASTM D86-17, Standard Test Method for Distillation of Petroleum Products and Liquid Fuels at Atmospheric Pressure, ASTM International, 2017, West Conshohocken, PA
37. Plehiers, P.P.; Symoens, S.H.; Amghizar, I.; Marin, G.B.; Stevens, C.V.; Van Geem, K.M. Artificial Intelligence in Steam Cracking Modeling: A Deep Learning Algorithm for Detailed Effluent Prediction. *Engineering* **2019,** *5*, 1027-1040.
38. König, A.; Marquardt, W.; Mitsos, A.; Viell, J.; Dahmen, M. Integrated design of renewable fuels and their production processes: recent advances and challenges. *Current Opinion in Chemical Engineering* **2020,** *27*, 45-50.
39. Dao Thi, H.; Djokic, M.R.; Van Geem, K.M. Detailed Group-Type Characterization of Plastic-Waste Pyrolysis Oils: By Comprehensive Two-Dimensional Gas Chromatography Including Linear, Branched, and Di-Olefins. *Separations* **2021,** *8*.
40. Kusenberg, M.; Zayoud, A.; Roosen, M.; Thi, H.D.; Abbas-Abadi, M.S.; Eschenbacher, A.; Kresovic, U.; De Meester, S.; Van Geem, K.M. A comprehensive experimental investigation of plastic waste pyrolysis oil quality and its dependence on the plastic waste composition. *Fuel Processing Technology* **2022,** *227*, 107090.
41. Walters, W.P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Accounts of Chemical Research* **2021,** *54*, 263-270.
42. Speight, J.G. *The Chemistry and Technology of Petroleum*. CRC Press, **2014**.
43. McKay, B.D.; Yirik, M.A.; Steinbeck, C. Surge - A Fast Open-Source Chemical Graph Generator. *ChemRxiv. Cambridge: Cambridge Open Engage; 2021; This content is a preprint and has not been peer-reviewed.* **2021** Version of 2021-12-08 DOI: 10.26434/chemrxiv-2021-gt5lb
44. *RDKit: Open-Source Cheminformatics*. rdkit.org, **2020**, accessed 2021-12-12
45. Yaws, C.L.; Narasimhan, P.K. In *Thermophysical Properties of Chemicals and Hydrocarbons*, Yaws, C.L.,Editor; William Andrew Publishing: Norwich, NY, **2009**; 1-95.
46. Yaws, C.L.; Pike, R.W. In *Thermophysical Properties of Chemicals and Hydrocarbons*, Yaws, C.L.,Editor; William Andrew Publishing: Norwich, NY, **2009**; 106-197.
47. Yaws, C.L.; Satyro, M.A. In *The Yaws Handbook of Vapor Pressure (Second Edition)*, Yaws, C.L.,Editor; Gulf Professional Publishing, **2015**; 1-314.
48. Daubert, T.E.; Danner, R.P. *Data compilation tables of properties of pure compounds*. Design Institute for Physical Property Data, American Institute of Chemical Engineers: New York, NY, **1985**.

49. Pyl, S.P.; Van Geem, K.M.; Reyniers, M.-F.; Marin, G.B. Molecular reconstruction of complex hydrocarbon mixtures: An application of principal component analysis. *AIChE Journal* **2010,** *56*, 3174-3188.
50. Mei, H.; Wang, Z.; Huang, B. Molecular-Based Bayesian Regression Model of Petroleum Fractions. *Industrial & Engineering Chemistry Research* **2017,** *56*, 14865-14872.
51. Riazi, M.R. *Characterization and properties of petroleum fractions*. ASTM international: West Conshohocken, PA, **2005**.
52. Peng, B., Molecular modelling of petroleum processes, *University of Manchester* **1999**
53. Mi Saine Aye, M.; Zhang, N. A novel methodology in transforming bulk properties of refining streams into molecular information. *Chemical Engineering Science* **2005,** *60*, 6702-6717.
54. Wu, Y.; Zhang, N. Molecular Characterization of Gasoline and Diesel Streams. *Industrial & Engineering Chemistry Research* **2010,** *49*, 12773-12782.
55. Gomez-Prado, J.; Zhang, N.; Theodoropoulos, C. Characterisation of heavy petroleum fractions using modified molecular-type homologous series (MTHS) representation. *Energy* **2008,** *33*, 974-987.
56. Ranzi, E.; Dente, M.; Goldaniga, A.; Bozzano, G.; Faravelli, T. Lumping procedures in detailed kinetic modeling of gasification, pyrolysis, partial oxidation and combustion of hydrocarbon mixtures. *Progress in Energy and Combustion Science* **2001,** *27*, 99-139.
57. Mei, H.; Cheng, H.; Wang, Z.; Li, J. Molecular characterization of petroleum fractions using state space representation and its application for predicting naphtha pyrolysis product distributions. *Chemical Engineering Science* **2017,** *164*, 81-89.
58. Ranzi, E.; Pierucci, S.; Dente, M.; Van Goethem, M.; Van Meeuwen, D.; Wagner, E. Correct molecular reconstruction of cracking feeds: a need for the accurate predictions of ethylene yields. *Chemical Engineering Transactions* **2015,** *43*, 871-876.
59. Mango, F.D. The light hydrocarbons in petroleum: a critical review. *Organic Geochemistry* **1997,** *26*, 417-440.
60. Mango Frank, D. An Invariance in the Isoheptanes of Petroleum. *Science* **1987,** *237*, 514-517.
61. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988,** *28*, 31-36.
62. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989,** *29*, 97-101.
63. Daylight Chemical Information Systems Inc. *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, **2019**, accessed 2021-12-10
64. Rossini, F.D. Hydrocarbons in petroleum. *Journal of Chemical Education* **1960,** *37*, 554.
65. Van Geem, K.M.; Hudebine, D.; Reyniers, M.F.; Wahl, F.; Verstraete, J.J.; Marin, G.B. Molecular reconstruction of naphtha steam cracking feedstocks based on commercial indices. *Computers & Chemical Engineering* **2007,** *31*, 1020-1034.
66. Bell, M.F. Analysis of East Texas Virgin Naphtha Fractions Boiling up to 270° F. *Analytical Chemistry* **1950,** *22*, 1005-1014.

67. Martin, R.L.; Winters, J.C. Determination of Hydrocarbons in Crude Oil by Capillary-Column Gas Chromatography. *Analytical Chemistry* **1963,** *35*, 1930-1933.
68. Ha, Z.; Ring, Z.; Liu, S. Estimation of Isomeric Distributions in Petroleum Fractions. *Energy & Fuels* **2005,** *19*, 1660-1672.
69. Riniker, S.; Landrum, G.A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015,** *55*, 2562-2574.
70. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** *17*, 490-519.
71. Chollet, F. *Keras*. https://keras.io, **2015**, accessed 2021-11-26
72. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. *Tensorflow: A system for large-scale machine learning*.
73. Glorot, X.; Bengio, Y., *Understanding the difficulty of training deep feedforward neural networks*. PMLR. p. 249-256.
74. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* **2014** Version of 2017-01-30 DOI: 10.48550/arXiv.1412.6980
75. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011,** *12*, 2825-2830.
76. Pitzer, K.S. The Volumetric and Thermodynamic Properties of Fluids. I. Theoretical Basis and Virial Coefficients1. *Journal of the American Chemical Society* **1955,** *77*, 3427-3433.
77. Shi, C.; Borchardt, T.B. JRgui: A Python Program of Joback and Reid Method. *ACS Omega* **2017,** *2*, 8682-8688.
78. Yaws, C.L. In *Transport Properties of Chemicals and Hydrocarbons (Second Edition)*, Yaws, C.L.,Editor; Gulf Publishing Company: Oxford, **2014**; 131-254.
79. Van Geem, K.M.; Reyniers, M.F.; Marin, G.B. Challenges of Modeling Steam Cracking of Heavy Feedstocks. *Oil & Gas Science and Technology - Rev. IFP* **2008,** *63*, 79-94.
80. Van Geem, K.M.; Pyl, S.P.; Reyniers, M.-F.; Vercammen, J.; Beens, J.; Marin, G.B. On-line analysis of complex hydrocarbon mixtures using comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* **2010,** *1217*, 6623-6633.
81. Bojkovic, A.; Dijkmans, T.; Dao Thi, H.; Djokic, M.; Van Geem, K.M. Molecular Reconstruction of Hydrocarbons and Sulfur-Containing Compounds in Atmospheric and Vacuum Gas Oils. *Energy & Fuels* **2021,** *35*, 5777-5788.
82. Bojkovic, A.; Vermeire, F.H.; Kuzmanović, M.; Dao Thi, H.; Van Geem, K.M. Analytics Driving Kinetics: Advanced Mass Spectrometric Characterization of Petroleum Products. *Energy & Fuels* **2022,** *36*, 6-59.

812 # TOC Graphic



813