AIChE

Molecular Reconstruction of Complex Hydrocarbon Mixtures: An Application of Principal Component Analysis

Steven P. Pyl, Kevin M. Van Geem, Marie-Françoise Reyniers, and Guy B. Marin
Laboratory for Chemical Technology, Ghent University, Krijgslaan 281 (S5), B-9000 Gent, Belgium

DOI 10.1002/aic.12224
Published online April 20, 2010 in Wiley Online Library (wileyonlinelibrary.com).

Three methods for reconstruction of the detailed molecular composition of complex hydrocarbon mixtures, based on their global properties, are compared: a method based on the Shannon entropy criterion, an artificial neural network and a multiple linear regression model. In spite of the broad range of naphthas included in the training set, the application range of the last two methods proved to be limited. Principal component analysis allowed to identify their three-dimensional ellipsoidal application range. In this subspace, the artificial neural network is more accurate than the multiple linear regression model and the Shannon entropy method. However, outside its application range, the performance of the neural network, as well as the regression model, decreases drastically. In contrast, the performance of the Shannon entropy method is not influenced by the characteristics of the considered naphtha, but rather depends on the number of available commercial indices. The Shannon entropy method yields comparable results to the artificial neural network, provided that a sufficient amount of distillation data is available to supply information on the carbon number distribution. Combining the reconstruction methods with a fundamental simulation model illustrates the necessity of having accurate feedstock reconstruction methods since they allow to capture the full power of fundamental simulation models for the simulation of industrial processes. © 2010 American Institute of Chemical Engineers *AIChE J,* 56: 3174–3188, 2010

Keywords: artificial neural network, molecular reconstruction, fundamental kinetic modeling, principal component analysis, process simulation, Shannon entropy

Introduction

Complex hydrocarbon mixtures like naphtha, kerosene, diesel, gas oil, etc., contain a multitude of components, including *n*-paraffins, *iso*-paraffins, olefins, naphthenes, and aromatics. When such mixtures are used as feedstock, reliable tools for determining a detailed molecular composition are crucial. Knowledge of an accurate feedstock composition, which, in industrial practice, can change on a daily or even hourly basis, makes it easier to estimate proper economical value, since it significantly affects product yields and quality.[1] Furthermore, an accurate molecular composition also allows to determine whether the feed meets the design and environmental specifications of the considered unit and to identify possible bottlenecks.

Knowledge of the detailed molecular feedstock composition is especially important when fundamental models are used for process simulation. Such models have been developed for various chemical processes, such as steam cracking and pyrolysis,[2–7] oxidation,[8–12] steam reforming,[13] hydrocracking,[14–19] catalytic cracking,[20–23] etc. These fundamental

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to M.-F. Reyniers at mariefrancoise.reyniers@ugent.be.

models are able to simulate the chemical kinetics over a wide range of process conditions and for a wide range of feedstock types, by accounting for the occurring chemical reactions as well as for the physical transport phenomena.[24] This allows to predict product yields and to determine optimal process conditions, i.e., provides a powerful tool to improve process performance. Figure 1 illustrates that the fundamental modeling goal is accomplished at the molecular level: the behavior of the feedstock molecules and the formation of the product molecules is described by a large network of elementary reactions and their associated kinetics. This, in contrast to so-called lumped models, where the feedstock molecules, the occurring reactions, and the product molecules are grouped together in so-called pseudocomponents to reduce complexity. Compared with fundamental models, these lumped models therefore lack fundamental kinetic information and flexibility in terms of feedstock and operating conditions.[25]

To obtain the detailed molecular feedstock composition, necessary to make optimal use of the fundamental process simulation strategy, several analytical techniques can be used, e.g., conventional gas chromatography (1D-GC) for lighter fractions, comprehensive two-dimensional gas chromatography (GC × GC) for heavier fractions,[26,27] etc. Although these techniques allow obtaining an accurate molecular composition, they are in general very time-consuming, and the obtained data is often difficult to interpret. Furthermore, for the heaviest petroleum fractions, e.g., heavy gas oil and vacuum gas oils, no current analytical technique is powerful enough to detect and quantify the thousands of different components, including a significant number of isomers, that make up such an oil fraction.[28]

In an effort to eliminate the use of time-consuming analytical techniques, current research focuses on the development of numerical methods that reconstruct the molecular composition of a mixture based on a number of average properties or so-called commercial indices.[29–42] Although they are not representative of all the chemical and structural variety that such mixtures can contain, process feedstock is generally identified and distinguished by these commercial indices rather than by its molecular composition.[28] These indices, e.g., the average molecular weight of the mixture, some points of a boiling point distillation curve, the specific density, the global PIONA weight fractions, etc., are determined by means of relatively simple and standardized analytical procedures.[43] Several global characterization methods have been developed based on gas chromatography (GC),[44] high-performance-liquid-chromatography (HPLC),[45] supercritical fluid chromatography (SFC),[46] mass spectroscopy (GC-MS),[47] comprehensive 2D GC (GC × GC) for complete characterization of complex mixtures[48] or for trace analysis, e.g., quantification of the sulfur amount,[49] etc. Also Fourier Transform-near infrared (FT-NIR)[50] and FT-nuclear magnetic resonance (FT-NMR),[51,52] can be used to determine global characteristics of petroleum fractions.

Until recently the importance of feedstock reconstruction was generally ignored in the fundamental modeling strategy. However, as illustrated in Figure 1, this step is crucial if fundamental models are to be applied on a regular basis to optimally operate and design chemical production units.

This article compares three methods for molecular reconstruction: maximization of the Shannon entropy, artificial neural networks, and multiple linear regression. Although these reconstruction methods can be applied to all kinds of mixtures, the emphasis in this work is on naphtha because its detailed composition can still be determined experimentally in reasonable time, especially when compared to, for example, heavy gas oil fractions. This allowed to obtain a large amount of experimental data, i.e., the so-called training set, necessary to develop the artificial neural network and the multiple linear regression model, as will be discussed below. The application range of an artificial neural network and a multiple linear regression model strongly depends on this dataset. Therefore, special attention will be paid to the unambiguous classification of petroleum fractions based on a principal component analysis of the training set. This classification is an important feature that allows to define the application range of the reconstruction methods and to determine a priori whether the considered feedstock falls herein.

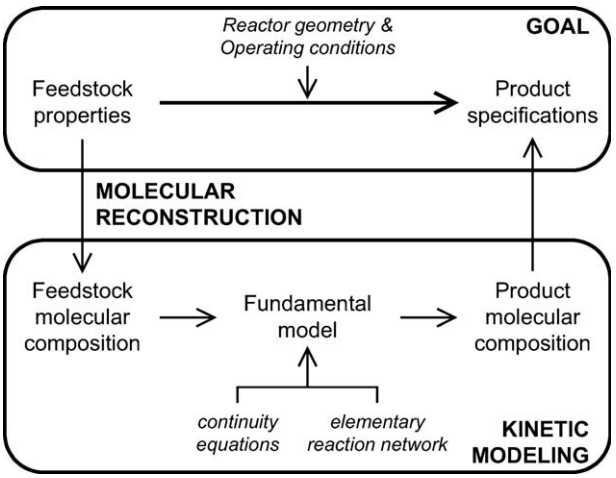

Figure 1. Fundamental kinetic modeling approach: from feed to product.[25]

Finally, the reconstruction methods are coupled to a fundamental simulation model for steam cracking, illustrating the use of the proposed simulation strategy and the evaluation of the influence of differences between a reconstructed and an analytically determined composition on the simulation results.

Reconstruction Methods

Starting from a number of average mixture properties, it is not straightforward to determine the corresponding molecular composition, simply because there is no unique relationship between a set of commercial indices and the detailed composition of a complex mixture. Molecular reconstruction therefore corresponds to the selection of a single molecular composition, i.e., the one most likely to occur, out of all theoretically possible molecular compositions that comply with the specified commercial indices.

A distinction can be made between two types of methods for molecular reconstruction. Methods of the first type allow to determine a detailed molecular composition by optimizing a specific objective function, subject to constraints that are derived from the available commercial indices. The objective

function can be derived from theoretical concepts such as Gibbs free energy[39] or Shannon entropy,[35,41] or can be some sort of cost function.[37,38] Several of these methods start from a predefined set of components, the mole fractions of which are adjusted using a specific optimization algorithm.[29,37,41] In other approaches, the algorithm that adjusts the mole fractions is preceded by an algorithm that generates the specific set of molecules to be considered. To create such a set of molecules, several possibilities exist, e.g., algorithms using group contribution methods[1,40] or stochastic methods.[30,31,35,36,39]

The second type of reconstruction methods uses a rather pragmatic approach, as they are based on a large set of experimental data, the so-called training set. Basically, the composition is reconstructed by interpolation between the samples in the training set using, for example, an artificial neural network[33,34] or other empirical correlations.[2,42] Generally, these reconstruction methods are faster than those of the first type since they are computationally less demanding. A disadvantage is that, because the size of the employed training set is evidently finite, the application range of these methods will be determined by the latter. These methods therefore require to determine a priori whether the considered feedstock falls within the application range defined by the training set. Frequently, nonphysicochemical criteria and/or expert user involvement are employed, making this classification error-prone, not transparent and non extendible. In this work, a principal component analysis (PCA) of the training set is used to define the application range and to compare petroleum fractions unambiguously.

In the following paragraphs three reconstruction methods are discussed. The first method belongs to the first type of methods and employs the so-called Shannon entropy criterion. The other two methods, based on artificial neural networks and multiple linear regression, belong to the second type of reconstruction methods. As will be discussed, the artificial neural network allows to model the relationship between commercial indices and detailed composition using a nonlinear relationship. While this is also possible using multiple linear regression, this work only considers a strictly linear regression model, i.e., linear in both model parameters and variables. The following paragraphs therefore allow the comparison of a linear approach and a nonlinear approach.

Shannon Entropy Maximization

The first reconstruction method discussed in this article follows a rather theoretical approach. When the maximization of the Shannon entropy (MSE), originally formulated in the information theory developed by Shannon,[53] is applied to the molecular reconstruction of hydrocarbon mixtures, the Shannon entropy criterion is defined by Eq. 1.[35,41]

$$\text{MAX} \quad S(y_i^{\text{rec}}) = -\sum_{i=1}^{N_m} y_i^{\text{rec}} \ln(y_i^{\text{rec}}) \quad \text{with} \quad \sum_{i=1}^{N_m} y_i^{\text{rec}} = 1 \quad (1)$$

S represents the Shannon entropy and y_i is the mole fraction of one of the N_m components that is part of the considered set of molecules. The principle of maximum Shannon entropy states that if only partial information concerning possible outcomes is available, the variables, i.e., the mole fractions, are to be chosen so as to maximize the uncertainty on the missing information, which is represented by the Shannon entropy S. This criterion can also be related to the thermodynamics of mixtures since the Shannon entropy, defined by Eq. 1, is directly proportional to the mixing entropy.[54] The Shannon criterion therefore corresponds to the assumption that mixtures with a composition that results in a higher entropy, and therefore lower Gibbs free energy, are favored in nature.

Evidently, the Shannon entropy has to be maximized subject to some constraints like those in Eq. 2. These constraints are derived from the N_p available commercial indices and express their dependency on the unknown mole fractions using mixing rules and other correlations.

$$p_k - f_k(y_i^{\text{rec}}) = 0 \quad k = 1, \ldots, N_p \quad (2)$$

This means that out of all the possible mixture compositions that meet the specified boundary conditions a single composition is selected, i.e., the composition with maximum Shannon entropy. In the absence of any information, it is impossible to favor one molecule over another, and the distribution of mole fractions will be completely uniform.

The so-called molecular library forms the basis of the method and specifies which molecules are considered. For each component, the library should contain all physical properties, e.g., normal boiling point, molecular weight, density, etc, necessary to derive the constraints in Eq. 2. Depending on the type of the considered mixture, e.g., naphtha or kerosene, a different molecular library has to be selected. Evidently, a library containing components that are not representative for the considered mixture will never result in an accurate reconstruction. Furthermore, if the reconstructed composition is used as input for a fundamental kinetic model, the library of components should be aligned with the required information on feedstock composition, as illustrated in the last paragraph, which means the library development also depends on the final application of the reconstructed composition. For the reconstruction of naphtha, a molecular library containing 118 different molecules was created. The experimental compositions of a number of well chosen naphthas, determined using one-dimensional gas chromatography, showed that these 118 molecules, including *n*-paraffins (C_3–C_{13}), *iso*-paraffins (C_4–C_{13}), olefins (C_4–C_7), naphthenes (C_5–C_{11}), and aromatics (C_6–C_{11}), are the components found in most petroleum naphthas. The complete list of components can be found in supporting information. All physical properties included in the library were taken from NIST Chemistry WebBook.[55]

Using this rather large library allows to reconstruct a much more detailed molecular composition as compared to the composition reconstructed with the library proposed by Van Geem et al.[41] which included only 37 chemical components. Constructing a relevant molecular library for heavier fractions, e.g., kerosene, can be carried out similarly but requires more advanced analytical techniques such as comprehensive GC $\times$ GC.[26] However, for the heaviest fractions, e.g., vacuum gas oil, no current analytical technique is powerful enough to characterize the complete mixture. In this case, stochastic methods can be used to come up with an appropriate library.[30,31,35]

Finding the global optimum of the nonlinear problem defined by Eqs. 1 and 2 is not straightforward, but one of

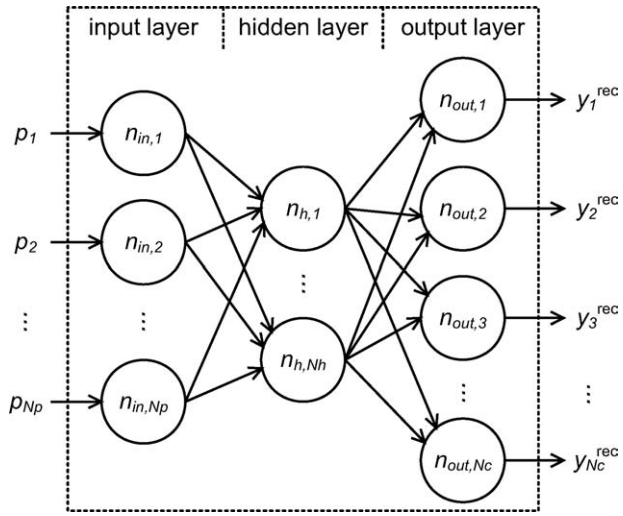

Figure 2. Schematic representation of an ANN for molecular reconstruction.

the advantages of the MSE method, compared to the other optimization methods, is the solution strategy for the maximization problem. The Lagrange multiplier method can be used to reduce the optimization problem with constraints to an optimization problem without constraints. Furthermore, if all constraints considered, see Eq. 2, are linear functions of the unknown mole fractions, the linearity of these constraints can be exploited to drastically simplify the objective function,[35,41] which is optimized using the Rosenbrock method.[56]

From all commercial indices considered in this work, i.e., average molecular weight, density, molar H/C ratio, global PIONA weight fractions, and up to 13 points from a true boiling point curve, such linear constraints can be derived. Additionally, more commonly used distillation data, e.g., ASTM D86 or ASTM D2887, can be converted, using Daubert's empirical conversion correlations,[43] into true boiling point data, since the latter is very rarely available in an industrial environment. Moreover, other types of compositional data can be incorporated into the reconstruction method, including SARA weight fractions, weight fractions of component groups identified by mass spectroscopy, ^{1}H- and ^{13}C-NMR data, etc., thus providing additional information on mixture composition.

Finally, from the reconstructed molecular composition and the information available from the molecular library, the commercial indices of the considered petroleum fraction are calculated using mixing rules and other correlations.[43] This allows to compare the characteristics of the generated mixture with the experimentally determined commercial indices and to determine whether the algorithm succeeded in finding a suitable optimum, which should correspond, besides to a maximal value of the Shannon entropy, to a minimal difference between the experimental commercial indices and the calculated ones.

Artificial Neural Networks

An artificial neural network (ANN) allows to develop complex and nonlinear relationships between multiple input and output variables. In recent years, this technique has gained wide popularity in many areas of chemical engineering.[57–59]

The concept of an ANN is loosely based on the human brain, which consists of over 10 billion neurons that are closely interconnected. When a neuron is activated by receiving a signal from a neighboring neuron it can fire off an electrochemical signal of its own. Artificial neural networks use this principle to model complex relationships. The input variables offered by the user are processed by the ANN, thus generating values for the output variables, which depend on network structure and connectivity. It is important to realize that such an ANN operates as a "black box" since the network parameters have no physical meaning. Figure 2 shows a schematic representation of a multilayer ANN that is used to determine the molecular composition of a hydrocarbon mixture, based on a number of commercial indices. The neurons are represented by circles. The weighted connections between them are unidirectional and are represented by arrows.

Neurons that receive input, i.e., values for the N_p selected commercial indices, are part of the input layer. These neurons perform no operation on the received information, but simply pass it on to the following neurons. For the molecular reconstruction of naphtha, the nine commercial indices given in Table 1 were selected as input variables, since they are commonly available in an industrial environment and contain information about the distribution in hydrocarbon classes, e.g., PIONA weight fractions, as well as the carbon number distribution, e.g., density and distillation data.

Neurons that are part of the output layer generate values for the mole fractions, y^{rec}, of the N_c chemical components considered in the model. For naphtha, the MSE method discussed in the previous section determines the mole fractions of 118 different molecules, commonly encountered in a wide range of naphtha fractions. However, constructing an ANN that generates an output of 118 mole fractions would be unachievable, as will be discussed below. Therefore, a partially lumped composition of 28 components is proposed, see Table 2. This composition that combines for example all *iso*-paraffins of a given carbon number into one so-called lumped component, contains sufficient information about the molecular composition of the naphtha. As previous studies have shown that the internal distribution of isomers is fairly independent from the feedstock,[60,7] this approach allows to determine a detailed molecular composition starting from the

Table 1. Ranges of the 9 Commercial Indices Included in the Naphtha Training Set

Commercial Index	Minimum Value	Maximum Value
Initial boiling point (IBP) [K]*	302.9	326.2
50 vol % boiling point ($T_{50\%}$) [K]*	325.7	389.7
Final boiling point (FBP) [K]*	353.2	453.2
Specific gravity (60-60°F) [−]	0.664	0.732
Total amount of *n*-paraffins [wt %]	28.5	49.8
Total amount of *iso*-paraffins [wt %]	27.3	51.7
Total amount of olefins [wt %]	0.00	1.00
Total amount of naphthenes [wt %]	5.86	33.3
Total amount of aromatics [wt %]	1.90	16.3

*Determined by ASTM D86 distillation.

Table 2. 28 Output Parameters of the ANN and MLR Model

n-Paraffins	*iso*-Paraffins	Olefins	Naphthenes	Aromatics
n-butane	C_5 *iso*-paraffins	C_5 olefins	Cyclopentane	Benzene
n-pentane	C_6 *iso*-paraffins	C_6 olefins	Cyclohexane	Ethylbenzene
n-hexane	C_7 *iso*-paraffins		Methylcyclopentane	Xylenes
n-heptane	C_8 *iso*-paraffins		C_7 naphthenes	C_8 aromatics
n-octane	C_9 *iso*-paraffins		C_8 naphthenes	C_9 aromatics
n-nonane	C_{10} *iso*-paraffins		C_9 naphthenes	
n-decane	C_{11} *iso*-paraffins			
n-undecane				

generated partially lumped composition, by assuming a fixed internal composition of the lumped components.

The nonlinear nature of the ANN is realized in two ways. First, besides input and output neurons a number of additional neurons, located in so-called hidden layers, are incorporated in the network, see Figure 2. Secondly, hidden neurons as well as output neurons perform an operation on the signal they receive. For example, a certain neuron receives n_s signals S_j from n_s preceding neurons. The total input signal $S_{\text{in},i}$ of this neuron is equal to the weighted sum of the received signals.

$$S_{\text{in},i} = \sum_{j=1}^{n_s} \omega_{i,j} \times S_j \tag{3}$$

where $\omega_{i,j}$ is the weight of the connection between the considered neuron and each of its preceding neighbors. The signal $S_{\text{in},i}$ is used as input for the activation function that characterizes the neuron. The result of this function, generally a sigmoid function like the one in Eq. 4,[61] is the output signal $S_{\text{out},i}$ of the neuron.

$$S_{\text{out},i} = \frac{1}{1 + e^{-(S_{\text{in},i} - \alpha_i)}} \tag{4}$$

where α_i is the so-called activation value of the neuron. The output of the neuron is passed onto the output neurons in case of a hidden neuron, or is equal to a final output value in case of an output neuron.

Determining the connectivity of the network, i.e., the activation values and the weights of the connections, is accomplished using specialized software (EasyNN Plus) that employs the backpropagation algorithm, thus minimizing the so-called training error (TE), defined by Eq. 5.

$$\text{MIN} \quad \text{TE} \propto \sum_{i=1}^{n} \sum_{j=1}^{N_c} \left(y_{i,j}^{\text{exp}} - y_{i,j}^{\text{rec}} \right)^2 \tag{5}$$

Since the output (y^{rec}) of the network also depends on the number of hidden neurons, the employed software is also able to determine the optimal number of hidden neurons in the hidden layer. According to the theory of Kolmogorov, an ANN with one hidden layer including $2N_p + 1$ hidden neurons is able to model the relationship between any specified set of input and output parameters.[62,57] Starting from this number of hidden neurons, i.e., 19 for the ANN considered in this work, the training software was set to determine the optimal number, i.e., 14 for the discussed ANN.

Equation 5 implies that training of an ANN requires an extensive training set containing n training samples, for which both the commercial indices as well as the associated values of the mole fractions are known from experimental work. The selection of adequate commercial indices, as well as appropriate output variables, prior to the method development is essential and strongly depends on the desired application of the generated chemical compositions. It is important that the number of training samples is large enough, to assure a sufficient degree of generalization. Since the required number of training samples increases dramatically with the number of parameters included in the network,[61] it was opted to restrict the output to a partially lumped composition of 28 components, as mentioned before. The training set used for the molecular reconstruction of naphtha fractions and some associated implications are discussed in greater detail further on.

Multiple Linear Regression

Multiple linear regression (MLR) can be used to model a relationship between a dependent variable and a number of independent variables. The latter are either control variables or are variables that can easily be observed. The objective of the regression model is to predict the dependent variable from the independent variables. The correlation between them is a linear combination of the model parameters, β[63].

Multiple linear regression can be used for molecular reconstruction, in which case the independent variables are a number of commercial indices while the mole fractions of the considered components in the mixture are the dependent variables. The latter implies that for each of the considered chemical components in the hydrocarbon mixture, a regression model has to be determined, as illustrated in Figure 3.

The ANN discussed earlier allowed to model the relationship between the mole fractions and the commercial indices using a nonlinear approach. To examine the possibility to accurately reconstruct a molecular composition using a strictly linear relationship, a regression model in the form of Eq. 6 is considered in this work.

$$y_j^{\text{rec}} = \beta_{0,j} + \beta_{1,j} \times p_1 + \ldots + \beta_{Np,j} \times p_{Np} \quad j = 1, \ldots, N_c \tag{6}$$

with N_p, the number of commercial indices and N_c, the number of chemical components considered in the model. This regression model does not include any quadratic terms or cross products, and is therefore linear in both model parameters and variables.

Concerning the molecular reconstruction of naphtha, a multiple linear regression model is developed starting from

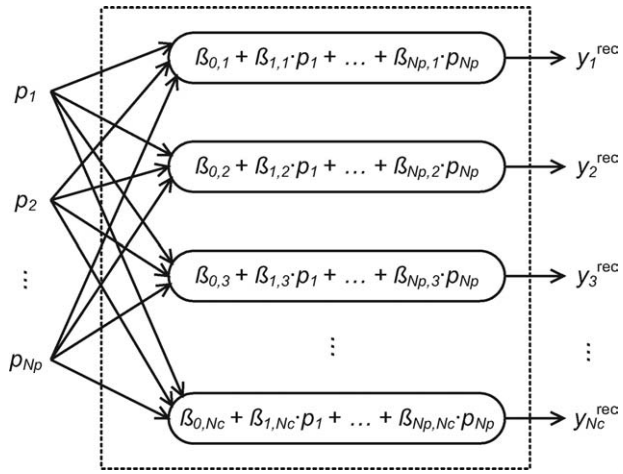

Figure 3. Schematic representation of a MLR model for molecular reconstruction, linear in both model parameters (β) and variables (p).

the same input and output parameters as those used for the ANN, Table 1 and Table 2. The regression model therefore includes 28 correlations in the form of Eq. 6, each of which are determined by 10 ($=N_p + 1$) regression coefficients $\beta_{i,j}$, see Figure 3. This implies that initially no detailed molecular composition is determined by the MLR model, but rather a partially lumped composition. The latter can be transformed to a detailed molecular composition by assuming a fixed internal composition of the lumped components,[60,7] see supra.

The regression coefficients of each of the 28 correlations are determined using the method of least squares, thus minimizing 28 separate objective functions, i.e., the sums of the squared errors (SSE) given in Eq. 7.[63]

$$\text{MIN} \quad SSE_j = \sum_{i=1}^{n} \left(y_{i,j}^{\text{exp}} - y_{i,j}^{\text{rec}} \right)^2 \quad j = 1, \ldots, N_c \qquad (7)$$

The least squares method requires an extensive and relevant training set containing n training samples, for which both the commercial indices as well as the associated values of the mole fractions are known from experimental work. This makes the concept of multiple linear regression rather similar to that of artificial neural networks, since both their development is based on an extensive training set containing experimental data.

Training Set

The training set, containing the required information for a large amount of industrial naphthas, was established over a period of almost 3 years. For all training samples nine commercial indices, specified in Table 1 with their corresponding ranges, are available. Besides these indices, the detailed molecular composition of each sample is also available. As discussed earlier, this molecular composition was reduced to a composition of 28 chemical components, given in Table 2, to facilitate the development of the ANN.

The requirement for a training set has several implications for the ANN as well as for the MLR model. First of all, the ANN and MLR model can only be applied if all of the commercial indices considered during the method development are available. If, for given naphtha, more commercial indices are available, they cannot be included in the reconstruction. If less commercial indices are available, the neural network or regression model cannot be used. This is an important difference compared with MSE, which can perform a reconstruction based on varying numbers and types of commercial indices.

Secondly, for heavier fractions such as kerosene or gas oil, it is not straightforward to gather even a few complete detailed compositions, implying that both ANN and MLR are not easily extended to heavier fractions due to a lack of training data.

Thirdly, the training set determines the application range of the reconstruction methods. It is obvious that these methods will not be able to provide accurate results for naphthas that differ substantially from the training samples. The parity diagram shown in Figure 4, where the analytical composition of an evaluation sample (naphtha A), i.e., a sample that is not part of the training set, is compared to its reconstructed composition, shows that MLR and ANN can be remarkably successful in predicting an accurate molecular composition. However, the parity diagrams for naphtha B and naphtha C, shown in Figure 5a and b, confirm the expectation that the application range of both reconstruction methods is limited. The commercial indices used for the molecular reconstruction of naphtha B indicate that this naphtha contains notably more olefins, i.e., 4.17 wt %, than the training naphthas, which contain between 0 and 1 wt % of olefins, see Table 1. This could explain why the reconstruction of this sample is so poor. On the other hand, although all the commercial

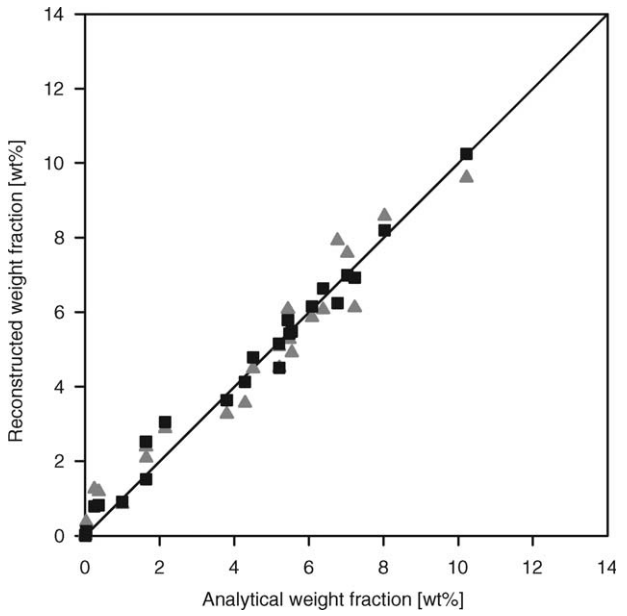

Figure 4. Parity diagram for naphtha A [d = 0.710 kg/m^3, P = 34.4 wt %, I = 28.1 wt %, O = 0.01 wt %, N = 27.4 wt %, A = 10.1 wt %, ASTM D86 IBP = 303.9 K, $T_{50\%}$ = 322.4 K, FBP = 419.7 K] (d_M = 1.9, MAD_{ANN} = 0.2, MAD_{MLR} = 0.4) (■ ANN, ▲ MLR).

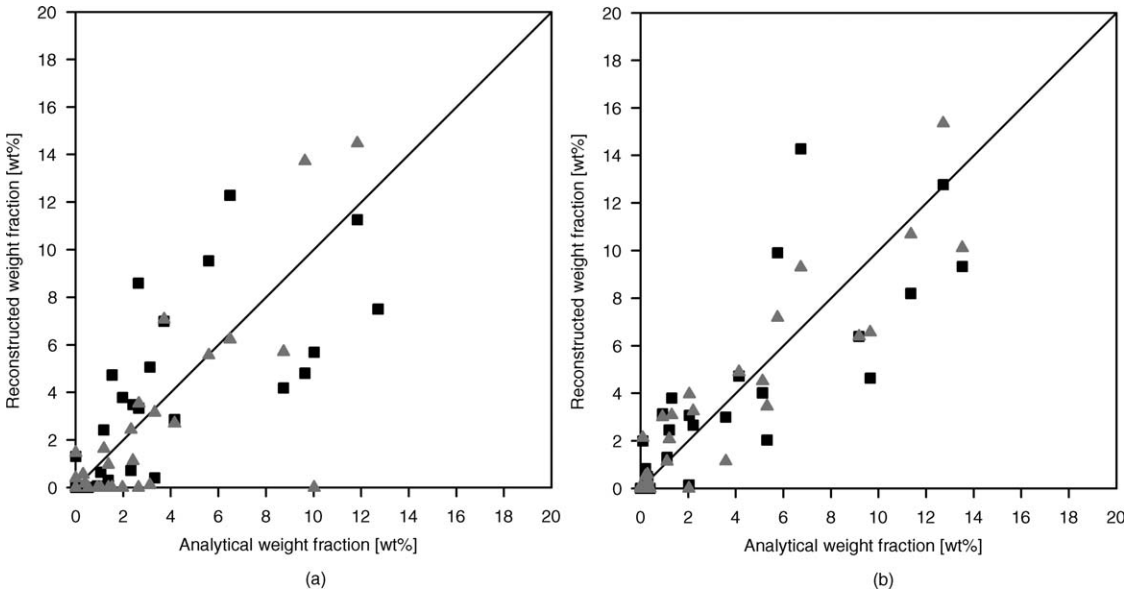

Figure 5. Parity diagram for (a) naphtha B [d = 0.693 kg/m^3, P = 29.9 wt %, I = 35.2 wt %, O = 4.17 wt %, N = 24.6 wt %, A = 6.13 wt %, ASTM D86 IBP = 306 K, $T_{50\%}$ = 341.6 K, FBP = 438.2 K] (d_M = 14, MAD_{ANN} = 2.3, MAD_{MLR} = 3.1) and (b) naphtha C [d = 0.690 kg/m^3, P = 37.7 wt %, I = 47.3 wt %, O = 0.2 wt %, N = 12.06 wt %, A = 2.71 wt %, ASTM D86 IBP = 308.5 K, $T_{50\%}$ = 362.5 K, FBP = 433.6 K] (d_M = 3.1, MAD_{ANN} = 1.6, MAD_{MLR} = 1.3) (■ ANN, ▲ MLR).

indices of naphtha C are within the ranges indicated in Table 1, the reconstruction of this sample is not good either. This example points out that to determine the application range of the MLR and ANN methods, it is not sufficient to compare the available commercial indices with the ranges given in Table 1. Therefore, practical implementation of the ANN and MLR model requires a method for assessing the similarity between the considered naphtha and the training set. In this work, feedstock similarity is assessed based on a principal component representation of both the feedstock to be reconstructed and the training samples. This unambiguous approach allows to determine a priori, based on the available commercial indices, whether the considered naphtha falls within the application range of ANN and MLR.

Principal component analysis (PCA) is a multivariate statistical technique that allows to project the information carried by the original variables, i.e., the commercial indices, onto an equal or smaller number of uncorrelated variables, while preserving the most important information.[64,65] PCA is mathematically defined as an orthogonal linear coordinate transformation such that the greatest variance by any projection of the data comes to lie on the first coordinate, i.e., the first principal component, the second greatest variance on the second coordinate, i.e., the second principal component, and so on. This approach allows to identify patterns and similarities in the data. Furthermore, PCA can be used for dimensionality reduction by retaining only those principal components that contribute most to the data set variance, thus facilitating the search for patterns, which can be hard to find in multidimensional data for which no graphical representation is possible.

This methodology is applied to determine a principal component representation of the naphtha samples in the training set used to develop the ANN and MLR model. Each of these n samples is characterized by nine commercial indices; see Table 1, which are brought together in $n \times 9$ matrix $\mathbf{P}$. To eliminate undesirable scale effects induced by the different units of these indices, the data in $\mathbf{P}$ is autoscaled,[64,65] resulting in an $n \times 9$ matrix $\tilde{\mathbf{P}}$ containing dimensionless variables. Next, the covariance matrix $\mathbf{C}$ of data matrix $\tilde{\mathbf{P}}$ is determined, using Eq. 8.

$$\mathbf{C} = \frac{1}{n-1}\tilde{\mathbf{P}}^{\mathrm{T}}\tilde{\mathbf{P}} \tag{8}$$

with

$$\tilde{p}_{i,j} = \frac{p_{i,j} - m_j}{s_j}$$

and

$$m_j = \frac{1}{n}\sum_{i=1}^{n} p_{i,j}$$

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n} \left(p_{i,j} - m_j\right)^2$$

Accordingly, the matrix $\mathbf{C}$ is a square symmetric 9×9 matrix. To reveal hidden structure in the original data, the goal is double: to minimize correlation between the variables and to differentiate between important and less important information carried by the data.

The off-diagonal elements of $\mathbf{C}$, i.e., the covariances, are a measure for the former. Therefore, to achieve the first goal, the data in $\tilde{\mathbf{P}}$ should be transformed to a coordinate system, which results in uncorrelated data, and, consequently, a diagonal covariance matrix ($\mathbf{L}$). To determine this space, the eigenvalues, λ_j, and corresponding eigenvectors, v_j, of the covariance matrix $\mathbf{C}$ are calculated. These

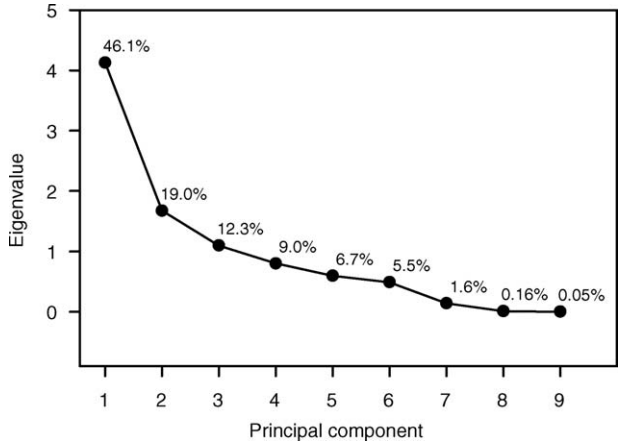

Figure 6. Eigenvalues and corresponding percentage of the total variance.

eigenvectors define the transformation between the original coordinate system and the coordinate system of principal components, while the eigenvalues are equal to the diagonal elements of $\mathbf{L}$ ($l_{j,j} = \lambda_j$), and thus correspond to the variances of the transformed data along each principal component.[64,65] The diagonal matrix $\mathbf{L}$ of eigenvalues and the matrix $\mathbf{V}$ of eigenvectors, which contains the eigenvector corresponding to the largest eigenvalue in the first column and so on, comply with Eq. 9.

$$\mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V} = \mathbf{L} \tag{9}$$

Consequently, the nine dimensional principal component representation of the data, in the form of an $n \times 9$ matrix $\mathbf{Z}$, is determined using Eq. 10.

$$\mathbf{Z} = \tilde{\mathbf{P}}\mathbf{V} \tag{10}$$

However, by considering only the principal components that correspond to the largest eigenvalues, i.e., restricting the matrix $\mathbf{V}$ to a smaller number of columns, the dimension of the transformed data is reduced, thus retaining only the most important information carried by the original data and ignoring so-called noise. According to the so-called *Scree* plot shown in Figure 6, which also shows the relative variances of the data along each principal component, the first three principal components account for nearly 80% of the total variance. Retaining only these three principal components allows to characterize every training naphtha using three variables instead of nine commercial indices. This 3D principal component representation ($\mathbf{Z}'$) is calculated by replacing $\mathbf{V}$ with $\mathbf{V}'$, i.e., a 9×3 matrix made up with the first 3 columns of $\mathbf{V}$, in Eq. 10. Retaining less than three principal components, results in the loss of important information about the data and an inadequate principal component representation.

The columns of $\mathbf{V}'$ contain, for each principal component, the so-called loadings corresponding to each of the nine (autoscaled) commercial indices. This information can be displayed in so-called loading diagrams. The interpretation of these diagrams is based on the direction in which the plotted values lie, as seen from the origin. Two properties are strongly correlated when there is a small angle between the lines connecting them with the origin.[64] Figure 7 shows such a loading diagram that allows to conclude that there is a strong correlation between the 50 vol % boiling point and the density and also between the initial and final boiling. The latter indicates that the training set primarily contains naphthas with similar boiling point range, i.e., the difference between FBP and IBP.

To determine whether a naphtha with unknown composition falls within the application range of the ANN and MLR model, requires an assessment of the similarity between the additional naphtha and the training naphthas, by comparing their principal component representations. The Mahalanobis distance (d_{M}), calculated using Eq. 11, is a measure for the distance between a single naphtha and the group of training naphthas that takes the variance of the data into account.[64]

$$d_{\mathrm{M}}^2 = \mathbf{z}^{\mathrm{T}} \times (\mathbf{L}')^{-1} \times \mathbf{z} \tag{11}$$

Where $\mathbf{z}$ is a column vector containing the three dimensional principal component representation of the naphtha, and $\mathbf{L}'$ is the 3×3 diagonal covariance matrix of the 3D principal component representation of the training set, and is therefore made up with the first 3 rows and columns of $\mathbf{L}$. If the Mahalanobis distance is small enough, the considered naphtha belongs to the group of training data and an accurate molecular reconstruction should be possible. Because the squared Mahalanobis distance follows the Hotelling T^2-distribution, the critical value can be obtained using Eq. 12.

$$\mathrm{T}_{\mathrm{crit}}^2 = \frac{3(n-1)}{n-3}\mathrm{F}_{3,n-3;0.05} = 2.5^2 \tag{12}$$

with $\mathrm{F}_{3,n\text{-}3;0.05}$, the F-statistic with (3,*n*-3) degrees of freedom corresponding to a three-dimensional ellipsoid that encloses 90% of the data. It is therefore assumed that naphthas with a principal component representation that falls within the

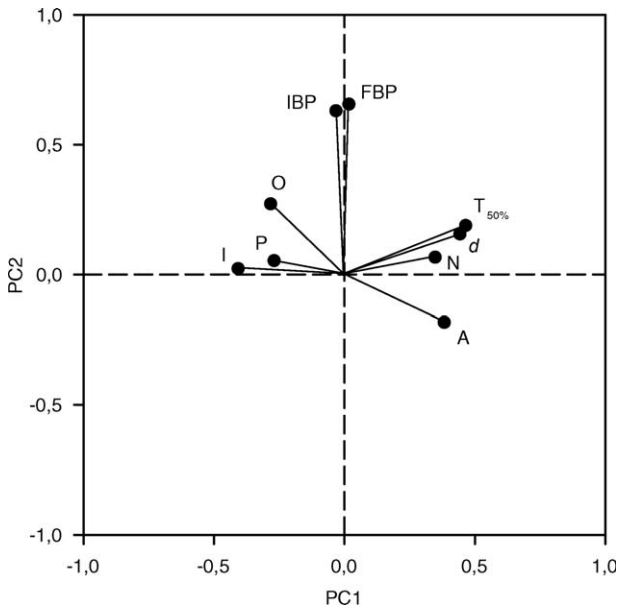

Figure 7. Loading diagram, loadings of PC1 vs. loadings of PC2.

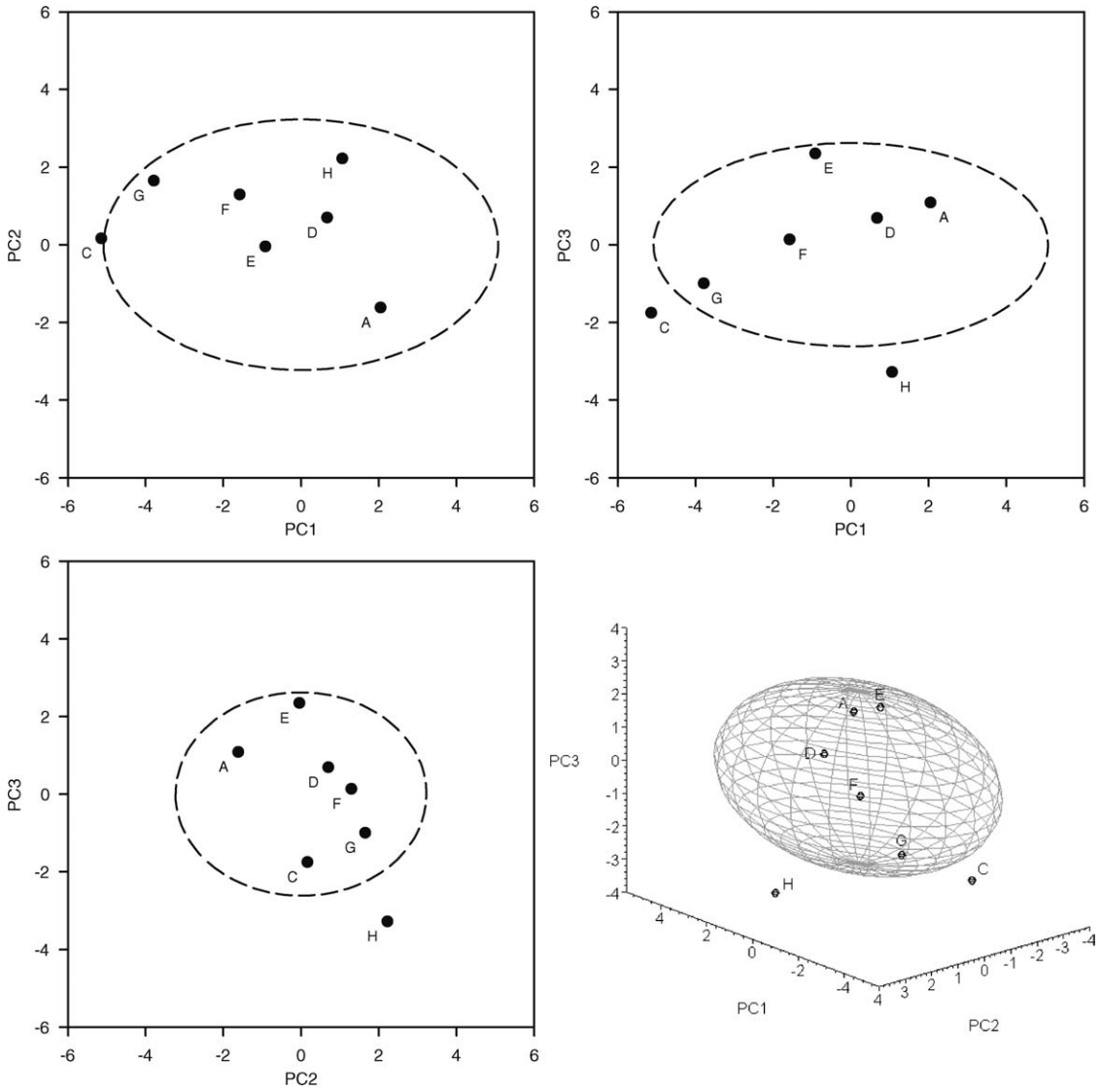

Figure 8. Three-dimensional principal component representation of naphtha A, C-H and ellipsoid corresponding to $d_M = 2.5$.

ellipsoid, i.e., naphthas for which this Mahalanobis distance is equal or smaller than 2.5, can be reconstructed accurately using the developed ANN or MLR model.

Figure 8 shows the principal component representation of seven of the eight naphthas discussed in this work along with the ellipsoid that corresponds to a Mahalanobis distance of 2.5. Naphtha B does not show up on these diagrams because the corresponding Mahalanobis distance simply is too large. From this figure it is clear that this approach is reliable, e.g., the Mahalanobis distance of the well reconstructed naphtha A shown in Figure 4 is smaller than 2.5, i.e., 1.9, whereas the Mahalanobis distance of naphtha B in Figure 5a as well as naphtha C in Figure 5b is larger than 2.5, i.e., 14.0 and 3.1, respectively.

Comparison of the Reconstruction Methods

The performance of the three methods for molecular reconstruction discussed earlier is compared in Table 3, which gives the average differences, i.e., the mean difference (MD), the mean absolute difference (MAD) and the root mean square difference (RMSD), between the reconstructed and the analytical weight fractions, per hydrocarbon class and per carbon number. All of the naphthas used to obtain the data in Table 3 fall within the application range of the ANN and MLR model.

It is obvious from both Table 3 and Figure 4 that the ANN and the MLR model are remarkably successful in predicting an accurate composition. Nevertheless, Table 3 clearly shows that the performance of the neural network is slightly better in most cases. The ANN seems to be better suited to model the complex relationship between the input and output parameters compared with the MLR model.

For naphthas within the application range of the ANN, the MSE method generally performs less accurate according to Table 3. However, Figure 9a shows that, in the case of naphtha D, the accuracy of MSE can be comparable to the accuracy of the ANN provided enough commercial indices are

Table 3. Average Differences* Between Analytical and Reconstructed Weight Fractions, Per Hydrocarbon Class and Per Carbon Number, for ñ Evaluation Naphthas

	MSE			ANN			MLR		
	MD [wt%]	MAD [wt%]	RMSD [wt%]	MD [wt%]	MAD [wt%]	RMSD [wt%]	MD [wt%]	MAD [wt%]	RMSD [wt%]
					n-Paraffins				
C_4	−1.43	1.75	2.21	0.27	0.50	0.68	0.03	0.52	0.68
C_5	1.68	2.28	2.72	0.02	0.97	1.35	−0.33	1.16	1.55
C_6	0.36	1.16	1.50	−0.15	0.71	1.03	0.04	1.10	1.43
C_7	−0.93	1.15	1.38	−0.25	0.47	0.64	0.09	0.63	0.95
C_8	−0.01	0.66	0.87	−0.08	0.29	0.37	0.00	0.50	0.64
C_9	−0.03	0.57	0.71	−0.11	0.26	0.35	−0.06	0.32	0.42
C_{10}	−0.22	0.27	0.43	−0.03	0.11	0.15	−0.06	0.22	0.27
C_{11}	0.02	0.05	0.08	0.01	0.03	0.04	−0.01	0.06	0.08
					iso-Paraffins				
C_5	2.04	2.40	3.04	0.08	0.82	1.08	−0.26	1.11	1.37
C_6	0.38	1.63	2.11	0.22	0.85	1.21	−0.18	1.40	1.82
C_7	−2.33	2.40	2.89	−0.52	0.84	1.11	0.02	1.02	1.41
C_8	−1.37	1.41	1.70	−0.29	0.44	0.57	−0.01	0.62	0.84
C_9	0.44	0.63	0.80	−0.09	0.32	0.43	−0.07	0.47	0.62
C_{10}	0.48	0.52	0.71	−0.06	0.25	0.32	−0.13	0.44	0.55
C_{11}	0.11	0.11	0.20	−0.01	0.04	0.06	−0.02	0.10	0.14
					Olefins				
C_5	0.00	0.01	0.03	−0.04	0.05	0.09	−0.04	0.04	0.07
C_6	0.02	0.02	0.04	−0.02	0.02	0.04	−0.03	0.03	0.05
					Naphthenes				
C_5	−2.20	2.20	2.32	0.00	0.14	0.19	−0.04	0.20	0.26
C_6	1.47	1.48	1.76	0.26	0.53	0.75	−0.29	1.07	1.35
C_7	2.16	2.18	2.38	−0.35	0.56	0.72	0.16	0.84	1.06
C_8	−0.34	0.56	0.74	−0.01	0.31	0.44	0.06	0.60	0.75
C_9	−0.90	0.93	1.16	−0.10	0.34	0.43	−0.07	0.46	0.64
					Aromatics				
C_6	−0.30	0.61	0.80	0.01	0.30	0.43	−0.17	0.54	0.76
C_7	0.79	0.81	1.00	−0.12	0.26	0.34	−0.03	0.45	0.60
C_8	−0.17	0.36	0.58	−0.11	0.26	0.35	−0.11	0.56	0.72
C_9	−0.12	0.58	0.69	−0.32	0.37	0.53	−0.22	0.38	0.49

* $\mathrm{MD} = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\left(x_i^{\mathrm{rec}} - x_i^{\mathrm{exp}}\right)$, $\mathrm{MAD} = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\left|x_i^{\mathrm{rec}} - x_i^{\mathrm{exp}}\right|$, $\mathrm{RMSD} = \sqrt{\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\left(x_i^{\mathrm{rec}} - x_i^{\mathrm{exp}}\right)^2}$

available, see Table 4. Especially the available distillation data has a considerable effect on the accuracy of the MSE method. These boiling points provide important information about the carbon number distribution in each hydrocarbon class (*n*-paraffins, *iso*-paraffins, olefins, naphthenes, or aromatics), thus reducing the uniform character of the distribution imposed by the objective function. A limited number of available commercial indices results in a more uniform distribution, which generally corresponds to an overestimation of components with the lowest or highest carbon numbers, e.g., C_5 and C_9 naphthalenes, and an underestimation of the middle ones, e.g., C_6 and C_7 naphthalenes. In contrast to Van Geem et al.,[41] no Gaussian distribution of the mole fractions as a function of the carbon number is imposed, and only the information contained within the available commercial indices is adopted to reconstruct the mixture composition.

Figure 9b shows, for naphtha E, the effect of the considered number of boiling points on the accuracy of the molecular reconstruction. It is obvious that the composition of naphtha E is reconstructed well, when all available distillation data, i.e., 13 boiling points, is employed. In this case, the MAD is equal to 0.6 wt %. Including less boiling points in the molecular reconstruction results in considerable deviations, especially when only three boiling points, i.e., IBP, $T_{50\%}$ and FBP, are used (MAD = 1.8 wt %). Using seven boiling points (IBP, $T_{10\%}$, $T_{30\%}$, $T_{50\%}$, $T_{70\%}$, $T_{90\%}$, FBP) results in a MAD of 0.8 wt %. The poor accuracy of the reconstructed composition in these cases can also be attributed to the used distillation data conversion methods. As mentioned earlier, to determine a molecular composition by MSE, the available distillation data (e.g., ASTM D86) has to be converted to a true boiling point curve. The accuracy of these conversion methods increases when more boiling points are available.[43] However, it must be emphasized that even when the reconstructed composition is not completely identical to the analytically determined one, the employed algorithm does succeed in finding an optimum for the considered objective function. This means that the generated composition will correspond to the specified commercial indices and to maximal Shannon entropy.

Finally, note that determining the application range is less of an issue when using the MSE method, in which case it is sufficient to create a suitable molecular library containing the components commonly found in a wide range of samples relevant for the desired application. Figure 10 shows that while the ANN is not able to accurately reconstruct the composition of naphtha C, a sample with a Mahalanobis distance larger than 2.5, the performance of the MSE method is not remarkably worse (MAD = 0.9) compared to for example the results obtained for naphtha E shown in Figure 9b (MAD = 0.7).

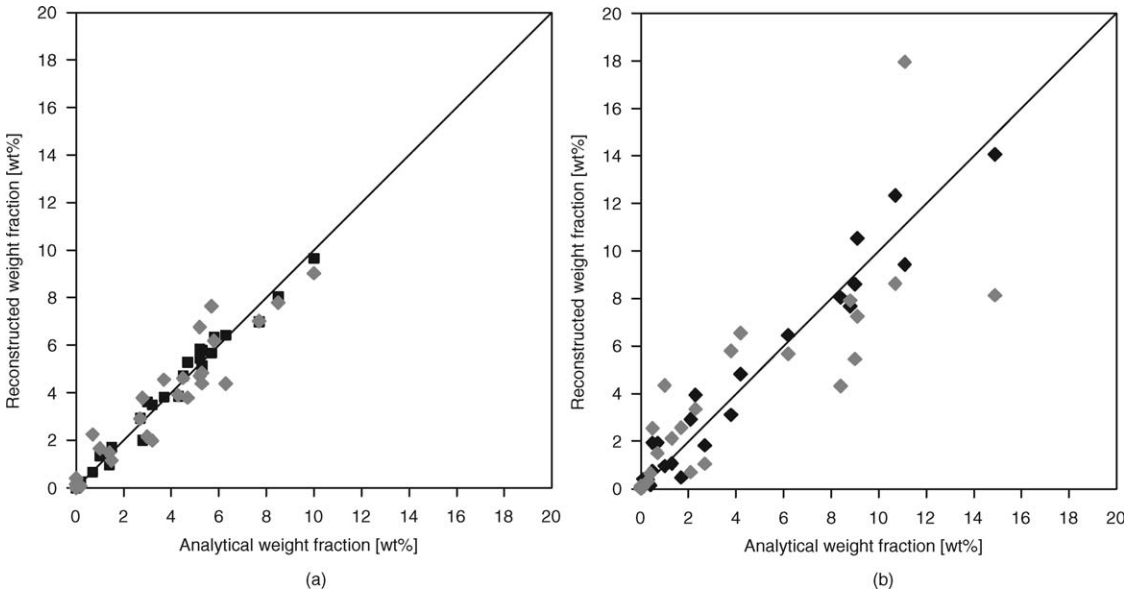

Figure 9. Parity plots for (a) naphtha D reconstructed with ANN (■) [d = 0.699 kg/m^3, P = 36.9 wt %, I = 32.9 wt %, O = 0.0 wt %, N = 21.4 wt %, A = 8.71 wt %, ASTM D86 IBP = 309.8 K, $T_{50\%}$ = 364.1 K, FBP = 440.6 K] and with MSE (◆) using all experimental commercial indices given in Table 4 and (d_M = 0.91, MAD_{ANN} = 0.3, MAD_{MSE} = 0.6) (b) naphtha E reconstructed with MSE using density, PIONA weight fractions and all boiling points given in Table 4 (◆, MAD = 0.7) and using density, PIONA weight fractions, IBP, $T_{50\%}$ and FBP given in Table 4 (◆, MAD = 1.8).

Validation of the Fundamental Kinetic Modeling Approach

The results in the previous paragraph show that the three considered methods are able to accurately reconstruct a feedstock composition based on the available commercial indices, provided a preliminary PCA assessment is performed. The effect of errors induced during reconstruction can be properly evaluated by combining the feedstock reconstruction methods with a fundamental kinetic model. In what follows, the reconstruction methods are combined with a fundamental model for steam cracking, COILSIM1D, that combines a single event microkinetic reaction network, CRACKSIM, with a set of one-dimensional reactor model equations that account for axial gradients in the reactor only.[7] This simplification is acceptable since, in this case, the model will be used to simulate the pilot plant installation for steam cracking of the Laboratory for Chemical Technology[66,67], which has a small enough diameter to neglect radial

Table 4. Experimental Commercial Indices and Those Corresponding to the Reconstructed MSE Composition, for Naphtha D, E, F, G and H

		Naphtha D		Naphtha E		Naphtha F		Naphtha G		Naphtha H	
		exp.	calc.	exp.	calc.	exp.	calc.	exp.	calc.	exp.	calc.
d	[kg/m^3]	0.699	0.697	0.691	0.693	0.687	0.687	0.679	0.678	0.714	0.711
P	[wt%]	37.0	37.9	38.6	38.8	34.2	34.2	36.3	36.5	25.7	25.2
I	[wt%]	32.9	33.8	32.1	32.0	35.8	35.8	42.0	41.7	32.3	32.5
O	[wt%]	0.00	0.00	0.00	0.00	0.10	0.10	0.20	0.20	0.95	0.96
N	[wt%]	21.4	20.5	22.8	22.7	26.6	26.7	17.3	17.5	34.5	34.9
A	[wt%]	8.71	7.87	6.61	6.56	3.23	3.25	4.17	4.19	6.50	6.44
IBP*	[K]	309.8	309.6	312.8	312.3	303.9	307.6	302.9	304.7	308.7	310.8
5%	[K]	320.5	318.6	321.3	321.4	315.1	317.1	312.0	311.9	323.7	322.0
10%	[K]	332.4	328.6	330.8	331.4	327.6	327.7	322.0	320.0	333.9	334.4
20%	[K]	339.5	337.2	337.9	338.9	332.6	333.4	325.8	324.1	346.9	346.5
30%	[K]	346.6	345.8	345.0	346.3	337.7	339.0	329.5	328.1	355.5	358.5
40%	[K]	355.3	354.8	350.3	352.3	341.2	342.1	333.9	332.8	365.2	365.7
50%	[K]	364.1	363.8	355.6	358.3	344.7	345.2	338.3	337.5	373.2	372.8
60%	[K]	376.5	374.1	359.7	361.8	352.7	352.2	348.5	347.6	381.3	380.7
70%	[K]	388.9	384.3	363.7	365.4	360.7	359.2	358.6	357.7	389.4	388.5
80%	[K]	399.6	397.4	373.9	376.3	378.1	376.7	381.9	381.1	399.0	400.8
90%	[K]	410.3	410.5	384.2	387.1	395.5	394.2	405.1	404.6	411.7	413.1
95%	[K]	426.2	426.5	401.0	404.0	411.6	410.5	423.6	422.8	420.5	421.9
FBP	[K]	440.6	441.0	416.0	419.2	426.0	425.2	440.2	439.2	431.8	429.8
d_M		0.91		2.28		1.27		2.45		3.60	

*All boiling points were determined by ASTM D86 distillation.

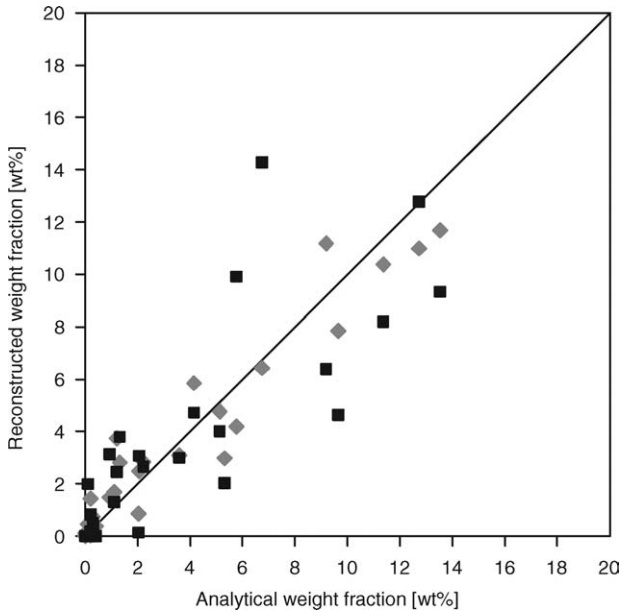

Figure 10. Parity diagram for *naphtha C* reconstructed with MSE (◆) [d = 0.690 kg/m^3, P = 37.7 wt %, I = 47.3 wt %, O = 0.2 wt %, N = 12.06 wt %, A = 2.71 wt %, ASTM D86 IBP = 308.5 K, $T_{5\%}$ = 319.8 K, $T_{10\%}$ = 332.2 K, $T_{20\%}$ = 339.4 K, $T_{30\%}$ = 346.7 K, $T_{40\%}$ = 354.6 K, $T_{50\%}$ = 362.5 K, $T_{60\%}$ = 366.8 K, $T_{70\%}$ = 371.1 K, $T_{80\%}$ = 384.9 K, $T_{90\%}$ = 398.6 K, $T_{95\%}$ = 417 K, FBP = 433.6 K] and with ANN (■) using only the 9 required commercial indices (d_M = 3.1, MAD_{ANN} = 1.7, MAD_{MSE} = 0.9).

nonuniformities.[68] As discussed by Van Geem et al.,[7] using CRACKSIM, requires a detailed molecular feedstock composition, since the model is made up with elementary reactions. The kinetic model considers over 478 components and distinguishes between different isomers in the naphtha range. Concerning naphtha cracking, the MSE molecular library, containing 118 components, is aligned with the required feedstock composition and compositions reconstructed with MSE are therefore sufficiently detailed for CRACKSIM. From the partially lumped composition generated with ANN or MLR the required detail can be derived by assuming a fixed internal composition of the lumped components, e.g., *iso*-paraffins.[60,7]

Pilot plant experiments with three naphthas are performed, i.e. naphtha F, G, and H, for which the process conditions were varied over a broad range. The flow rate of the hydrocarbon feedstock was varied between 9.9×10^{-4} kg/s and 1.1×10^{-3} kg/s, while the coil outlet temperature was varied from 819 K to 892 K. The dilution was varied from 0.40 $kg_{steam}/kg_{naphtha}$ to 0.70 $kg_{steam}/kg_{naphtha}$. The coil outlet pressure was varied from 0.15 to 0.20 MPa.

The detailed molecular composition of the naphthas is determined analytically using a gas chromatograph with a PONA column (50 m $\times$ 0.25 mm, 0.5 μm film). Furthermore, the feedstock composition is reconstructed using the available commercial indices of the naphthas (Table 4). The Mahalanobis distances given in Table 4 show that naphtha F is located inside the application range of the ANN, naphtha G is near the edge, and naphtha H is outside the ellipsoidal subspace.

The parity diagram in Figure 11a shows the yields of two important cracking products, i.e. ethene and propene, obtained when cracking naphtha F. This diagram allows to compare the experimental yields with the yields simulated with a reconstructed feedstock composition, obtained with either the MSE method or the ANN, for a naphtha within

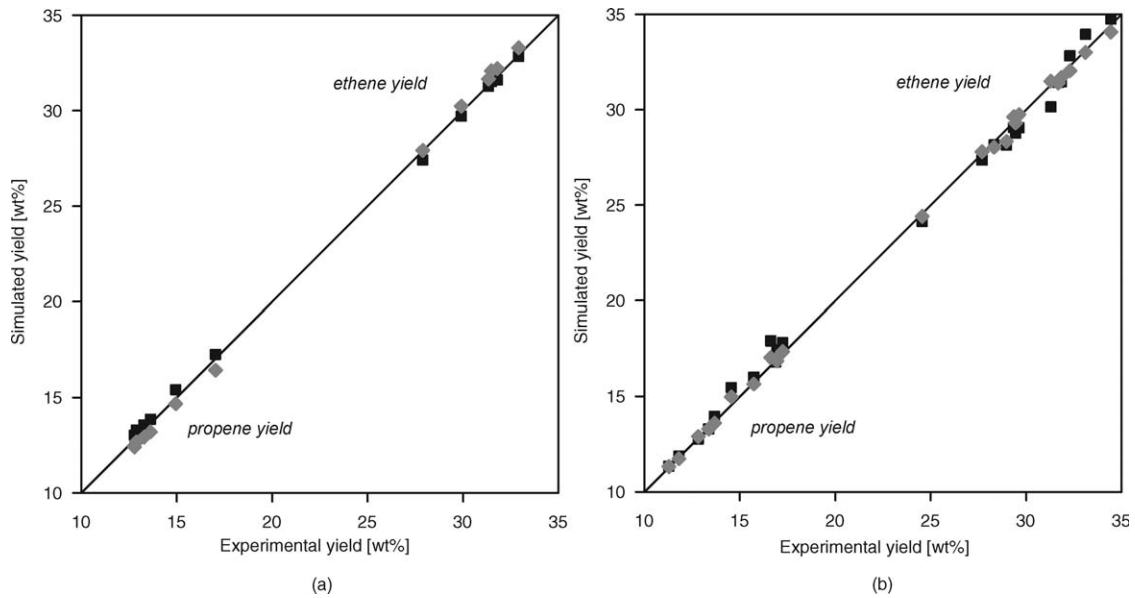

Figure 11. Pilot plant yields of ethene and propene, simulated with ANN composition (■) and MSE composition (◆), vs. the experimental yields; for (a) naphtha F and (b) naphtha G [ranges of process conditions: F = 9.9×10^{-4}–1.1×10^{-3} kg/s, COT = 819–892 K, δ = 0.40–0.70 kg/kg, COP = 0.15–0.20 MPa].

Table 5. Simulated and Experimentally Determined Pilot Plant Product Yields Obtained with *naphtha H* [Process Conditions: CIT = 873 K, COT = 1143 K, CIP = 0.23 MPa, COP = 0.17 MPa, $F = 1.2 \times 10^{-3}$ kg/s, $\delta = 1.0$ kg/kg, $\theta = 0.28$ s]

	Simulated yields [wt %]			Experimental yields [wt %]
	ANN	MSE	Analytical	
Hydrogen	0.71	0.72	0.74	0.97
Methane	15.15	14.58	14.32	14.09
Ethene	28.15	29.1	29.50	29.23
Ethane	3.91	3.69	3.57	3.44
Propene	16.83	17.59	17.71	17.63
1,3-Butadiene	3.93	4.37	4.69	4.71
1-Butene	2.11	2.32	2.23	2.13
2-Butene	0.94	0.81	0.99	1.31
Isobutene	4.32	3.51	3.40	3.67
Benzene	3.81	3.92	3.74	3.75
Toluene	1.48	1.39	1.16	1.26
Styrene	0.36	0.19	0.27	0.35
P/E [wt %/wt %]	0.60	0.60	0.60	0.60

the application range of the ANN. The results obtained with the reconstructed composition using MSE are similar to those obtained using ANN. For naphtha G, on the other hand, the results obtained with the ANN start to deviate slightly from the experimental yields, see Figure 11b. Naphtha G is on the border of the application range of the ANN, indicating that the performance of the ANN will not be perfect. Table 5 shows that for naphtha H the combination of reactor model and feedstock reconstruction using the ANN fails completely. Only for the analytically determined feed composition and the reconstructed feed composition using the MSE method, a proper agreement with the experimental data is obtained. The differences between the simulations using the ANN and the experimental data are even more important for some minor products such as ethane, iso-butene or 1,3-butadiene.

The results obtained for these three naphthas indicate that the feedstock composition has a considerable effect on the results of the simulation, showing that accurate analytical techniques or reliable molecular reconstruction methods are indispensible when using such fundamental simulation models. Finally, these results also illustrate the validity of the fundamental simulation strategy shown in Figure 1: the combination of an accurate reconstruction method with the fundamental simulation model allows to obtain accurate simulation results over a wide range of process conditions and feedstock compositions.

Conclusions

Three methods to reconstruct the detailed molecular composition of a hydrocarbon mixture, based on a number of commercial indices are compared: a method based on the Shannon entropy criterion, an artificial neural network and a multiple linear regression model.

The last two methods are both able to reconstruct the composition of naphtha fractions with great accuracy, provided the considered naphtha has similar characteristics compared to the large number of training naphthas used to develop these reconstruction methods.

A principal component analysis of the training set allowed to represent the training samples in a three-dimensional space and to identify unambiguously the application range of both methods. If the 3D principal component representation of a naphtha results in a Mahalanobis distance lower than 2.5, the naphtha falls within the application range. When this is not the case, significant deviations between the reconstructed and experimentally determined composition are observed.

In almost all cases, the overall performance of the ANN is better than the performance of the MLR model considered in this work. The strong nonlinear nature of the ANN seems better suited to model the complex relationship between the commercial indices and the considered chemical components compared to the linear relationship imposed by the MLR model considered in this work.

Within the application range defined by the training set, the ANN method is more accurate then the MSE method. However, outside this range the performance of the ANN decreases drastically, while the performance of the MSE is not influenced by the characteristics of the considered naphtha. The performance of the latter strongly depends on the number of available commercial indices and, in particular on the level of detail of the available distillation data. The main advantage of using MSE, is its ability of generating a feedstock composition based on varying numbers and types of commercial indices. Much information can be derived from a detailed PIONA analysis, which contains information about the distribution in hydrocarbon classes, and the density/distillation data, providing information about the carbon number distribution. However, the adoption of other types of feedstock properties, such as ^{1}H- and ^{13}C-NMR data, in the MSE method can yield additional compositional information and result in a even more detailed feedstock composition.

Finally, the three reconstruction methods were evaluated by combining them with a fundamental simulation model for steam cracking. Experimental product yields, obtained with a steam cracking pilot plant, were compared with yields that were simulated using a reconstructed feedstock composition. The results showed that a feedstock composition reconstructed with MSE as well as ANN, when applied inside its application range, enables accurate process simulation.

Notation

A = total amount of aromatics [wt %]
ANN = artificial neural network
CIP = coil inlet pressure [MPa]
CIT = coil inlet temperature [K]
COP = coil outlet pressure [MPa]
COT = coil outlet temperature [K]
d = density [kg/m^3]
d_M = Mahalanobis distance [-]
F = hydrocarbon mass flow rate [kg/s]
FBP = final boiling point [K]
I = total amount of *iso*-paraffins [wt %]
IBP = initial boiling point [K]
MD = mean difference, [wt %]
MAD = mean absolute difference [wt %]
MLR = multiple linear regression
MSE = maximization of the Shannon entropy
n = number of samples in the training set for ANN and MLR
N = total amount of naphthenes [wt %]
N_c = number of chemical components included in ANN and MLR model

N_h = number of hidden neurons in the ANN
N_m = number of molecules in the molecular library of the MSE method
N_p = number of commercial indices included in the ANN and MLR model
O = total amount of olefins [wt %]
P = total amount of n-paraffins [wt %]
PC = principal component
PCA = principal component analysis
RMSD = root mean square difference [wt %]
$T_{50\%}$ = 50 vol % boiling point [K]
x^{exp} = analytical weight fraction [wt %]
x^{rec} = reconstructed weight fraction [wt %]
y^{exp} = analytical mole fraction [mol %]
y^{rec} = reconstructed mole fraction [mol %]
δ = dilution [kg/kg]
θ = residence time [s]

Literature Cited

1. Quann RJ, Jaffe SB. Building useful models of complex reaction systems in petroleum refining. *Chem Eng Sci*. 1996;51:1615–1635.
2. Dente M, Ranzi E, Goossens AG. Detailed prediction of olefin yields from hydrocarbon pyrolysis through a fundamental simulation program SPYRO. *Comput Chem Eng*. 1979;3:61–75.
3. Chinnick SJ, Baulch DL, Ayscough PB. An expert system for hydrocarbon pyrolysis. *Chemom Intell Lab Syst*. 1988;5:39–52.
4. Hillewaert LP, Dierickx JL, Froment GF. Computer-generation of reaction schemes and rate-equations for thermal-cracking. *AIChE J*. 1988;34:17–24.
5. Broadbelt LJ, Stark SM, Klein MT. Computer-generated pyrolysis modeling—on-the-fly generation of species, reactions, and rates. *Ind Eng Chem Res*. 1994;33:790–799.
6. Matheu DM, Dean AM, Grenda JM, Green WH. Mechanism generation with integrated pressure dependence: a new model for methane pyrolysis. *J Phys Chem A*. 2003;107:8552–8565.
7. Van Geem KM, Reyniers MF, Marin GB. Challenges of modeling steam cracking of heavy feedstocks. *Oil Gas Sci Technol-Revue De L' Institut Francais Du Petrole*. 2008;63:79–94.
8. Chevalier C, Warnatz J, Melenk H. Automatic generation of reaction-mechanism for the oxidation of higher hydrocarbons. *Berichte Der Bunsen-Gesellschaft-Phys Chem Chem Phys*. 1990;94:1362–1367.
9. Dimaio FP, Lignola PG. KING, a kinetic network generator. *Chem Eng Sci*. 1992;47:2713–2718.
10. Blurock ES. Reaction—system for modeling chemical reactions. *J Chem Inform Comput Sci*. 1995;35:607–616.
11. Ranzi E, Faravelli T, Gaffuri P, Sogaro A. Low temperature combustion—automatic-generation of oxidation reactions and lumping procedures. *Combust Flame*. 1995;102:179–192.
12. Warth V, Battin-Leclerc F, Fournet R, Glaude PA, Come GM, Scacchi G. Computer based generation of reaction mechanisms for gas-phase oxidation. *Comput Chem*. 2000;24:541–560.
13. Maestri M, Vlachos DG, Beretta A, Groppi G, Rronconi E. A C-1 microkinetic model for methane conversion to syngas on Rh/Al2O3. *AIChE J*. 2009;55:993–1008.
14. Qader SA, Hill GR. Hydrocracking of gas oils. *Ind Eng Chem Process Des Dev*. 1969;8:98.
15. Baltanas MA, Vanraemdonck KK, Froment GF, Mohedas SR. Fundamental kinetic modeling of hydroisomerization and hydrocracking on noble metal loaded faujasites -1. Rate parameters for hydroisomerization. *Ind Eng Chem Res*. 1989;28:899–910.
16. Liguras DK, Allen DT. Comparison of lumped and molecular modeling of hydropyrolysis. *Ind Eng Chem Res*. 1992;31:45–53.
17. Quann RJ, Jaffe SB. Structure-oriented lumping—describing the chemistry of complex hydrocarbon mixtures. *Ind Eng Chem Res*. 1992;31:2483–2497.
18. Laxminarasimhan CS, Verma RP, Ramachandran PA. Continuous lumping model for simulation of hydrocracking. *AIChE J*. 1996;42:2645–2653.
19. Martens GG, Thybaut JW, Marin GB. Single-event rate parameters for the hydrocracking of cycloalkanes on Pt/US-Y zeolites. *Ind Eng Chem Res*. 2001;40:1832–1844.
20. John TM, Wojciechowski BW. Effect of reaction temperature on product distribution in catalytic cracing of neutral distillate. *J Catalysis*. 1975;37:348–357.
21. Jacob SM, Gross B, Voltz SE, Weekman VW. Lumping and reaction scheme for catalytic cracking. *AIChE J*. 1976;22:701–713.
22. Feng W, Vynckier E, Froment GF. Single-event kinetics of catalytic cracking. *Ind Eng Chem Res*. 1993;32:2997–3005.
23. Beirnaert HC, Alleman JR, Marin GB. A fundamental kinetic model for the catalytic cracking of alkanes on a USY zeolite in the presence of coke formation. *Ind Eng Chem Res*. 2001;40:1337–1347.
24. Froment GF. Kinetics and reactor design in the thermal-cracking for olefins production. *Chem Eng Sci*. 1992;47:2163–2177.
25. Klein MT, Hou G, Bertolacini RJ, Broadbelt LJ, Kumar A. *Molecular Modeling in Heavy Hydrocarbon Conversions*. Boca Raton: Taylor & Francis Group, 2006.
26. Phillips JB, Beens J. Comprehensive two-dimensional gas chromatography: a hyphenated method with strong coupling between the two dimensions. *J Chromatogr A*. 1999;856:331–347.
27. Vendeuvre C, Bertoncini F, Duval L, Duplan JL, Thiebaut D, Hennion MC. Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples. *J Chromatogr A*. 2004;1056:155–162.
28. Merdrignac I, Espinat D. Physicochemical characterization of petroleum fractions: the state of the art. *Oil Gas Sci Technol-Revue De L' Institut Francais Du Petrole*. 2007;62:7–32.
29. Allen DT, Liguras D. Structural models of catalytic cracking chemistry—a case study of group contribution approach to lumped kinetic modeling. *Chem React Complex Mixtures*. 1991;101–125.
30. Neurock M, Nigam A, Trauth D, Klein MT. Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms. *Chem Eng Sci*. 1994;49:4153–4177.
31. Campbell DM, Klein MT. Construction of a molecular representation of a complex feedstock by Monte Carlo and quadrature methods. *Appl Catalysis a-General*. 1997;160:41–54.
32. Bozzano G, Dente M, Sugaya M, McGreavy C. The characterization of residual hydrocarbon fractions with model compounds. Retaining the essential information. *Abstr Pap Am Chem Soc*. 1998;216:U869–U869.
33. Joo E, Park S, Lee M. Pyrolysis reaction mechanism for industrial naphtha cracking furnaces. *Ind Eng Chem Res*. 2001;40:2409–2415.
34. Lopez R, Perez JR, Dassori CG, Ranson A. Artificial neural networks applied to the operation of VGO hydrotreaters. In SPE Latin American Caribbean Petroleum Engineering Conference. 2001.
35. Hudebine D, Verstraete JJ. Molecular reconstruction of LCO gasoils from overall petroleum analyses. *Chem Eng Sci*. 2004;59:4755–4763.
36. Sheremata JM, Gray MR, Dettman HD, McCaffrey WC. Quantitative molecular representation and sequential optimization of Athabasca asphaltenes. *Energy Fuels*. 2004;18:1377–1384.
37. Albahri TA. Molecularly explicit characterization model (MECM) for light petroleum fractions. *Ind Eng Chem Res*. 2005;44:9286–9298.
38. Androulakis IP, Weisel MD, Hsu CS, Qian KN, Green LA, Farrell JT, Nakakita K. An integrated approach for creating model diesel fuels. *Energy Fuels*. 2005;19:111–119.
39. Ha ZY, Ring Z, Liu SJ. Derivation of molecular representations of middle distillates. *Energy Fuels*. 2005;19:2378–2393.
40. Jaffe SB, Freund H, Olmstead WN. Extension of structure-oriented lumping to vacuum residua. *Ind Eng Chem Res*. 2005;44:9840–9852.
41. Van Geem KM, Hudebine D, Reyniers MF, Wahl F, Verstraete JJ, Marin GB. Molecular reconstruction of naphtha steam cracking feedstocks based on commercial indices. *Comput Chem Eng*. 2007;31:1020–1034.
42. Ha HZ, Ring Z, Liu S. Data reconciliation among PIONA, GC-FIMS, and SimDis measurements for petroleum fractions. *Pet Sci Technol*. 2008;26:7–28.
43. Riazi MR. *Characterization and Properties of Petroleum Fractions*. West Conshohocken: ASTM International, 2005.
44. Beens J, Brinkman UAT. The role of gas chromatography in compositional analyses in the petroleum industry. *Trac-Trends Anal Chem*. 2000;19:260–275.
45. Kaminski M, Kartanowicz R, Gilgenast E, Namiesnik J. High-performance liquid chromatography in group-type separation and

technical or process analytics of petroleum products. *Crit Rev Anal Chem.* 2005;35:193–216.
46. Venter A, Rohwer ER. Comprehensive two-dimensional supercritical fluid and gas chromatography with independent fast programmed heating of the gas chromatographic column. *Anal Chem.* 2004;76:3699–3706.
47. Qian KN, Dechert GJ. Recent advances in petroleum characterization by GC field ionization time-of-flight high-resolution mass spectrometry. *Anal Chem.* 2002;74:3977–3983.
48. von Muhlen C, Zini CA, Caramao EB, Marriott PJ. Applications of comprehensive two-dimensional gas chromatography to the characterization of petrochemical and related samples. *J Chromatogr A.* 2006;1105:39–50.
49. Hua RX, Li YY, Liu W, Zheng JC, Wei HB, Wang JH, Lu X, Kong HW, Xu GW. Determination of sulfur-containing compounds in diesel oils by comprehensive two-dimensional gas chromatography with a sulfur chemiluminescence detector. *J Chromatogr A.* 2003;1019:101–109.
50. Cramer JA, Morris RE, Hammond MH, Rose-Pehrsson SL. Ultra-low sulfur diesel classification with near-infrared spectroscopy and partial least squares. *Energy Fuels.* 2009;23:1132–1133.
51. Bansal V, Vatsala S, Kapur GS, Basu B, Sarpal AS. Hydrocarbon-type analysis of middle distillates by mass spectrometry and NMR spectroscopy techniques—a comparison. *Energy Fuels.* 2004;18: 1505–1511.
52. Lough V. Consider MRA for online analysis—this new technique uses radio frequencies to do quantitative evaluations on hydrocarbon streams. *Hydrocarbon Process.* 2004;83:69–73.
53. Shannon CE. The mathematical theory of communication. *Bell System Tech J.* 1948;27:379–423, 623–656.
54. McQuarrie DA, Simon JD. *Molecular Thermodynamics.* Sausalito: University Science Books, 1999.
55. Linstrom PJ, Mallard WG. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Gaithersburg, MD: National Institute of Standards and Technology, 2009.
56. Rosenbrock HH. An automatic method for finding the greatest or least value of a function. *Comput J.* 1960;3:175–184.
57. Molga EJ. Neural network approach to support modeling of chemical reactors: problems, resolutions, criteria of application. *Chem Eng Process.* 2003;42:675–695.
58. Nabavi R, Niaei A, Salari D, Towfighi J. Modeling of thermal cracking of LPG: application of artificial neural network in prediction of the main product yields. *J Anal Appl Pyrolysis.* 2007;80:175–181.
59. Himmelblau DM. Accounts of experiences in the application of artificial neural networks in chemical engineering. *Ind Eng Chem Res.* 2008;47:5782–5796.
60. 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. *Prog Energy Combust Sci.* 2001;27:99–139.
61. Graupe D. *Principals of Artificial Neural Networks*. Singapore: World Scientific Publishing Co., 1999.
62. Kolmogorov AN. The representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk Sssr.* 1957;114:953–956.
63. Draper NR, Smith H. *Applied Regression Analysis*. New York: John Wiley & Sons, 1998.
64. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics, Part A & B*. Amsterdam: Elsevier, 1997, 1998.
65. Jolliffe IT. *Principal Component Analysis*. New York: Springer, 2002.
66. Vandamme PS, Froment GF. Putting computers to work—Thermal cracking computer control in pilot plants. *Chem Eng Prog*. 1982;78:77–82.
67. Van Geem KM, Reyniers MF, Marin GB. Two severity indices for scale-up of steam cracking coils. *Ind Eng Chem Res.* 2005;44:3402–3411.
68. Van Geem KM, Heynderickx GJ, Marin GB. Effect of radial temperature profiles on yields in steam cracking. *AIChE J.* 2004;50:173–183.

Manuscript received Aug. 28, 2009, and revision received Jan. 22, 2010.