ARTIFICIAL INTELLIGENCE IN STEAM CRACKING MODELING: DEEP LEARNING ALGORITHM FOR COMPOSITIONAL MODELING AND DETAILED EFFLUENT PREDICTION

Pieter P. Plehiers[1], Steffen H. Symoens[1], Ismaël Amghizar[1], Guy B. Marin[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 914 9052 Gent, Belgium

[2] Syn BioC Research Group, Department of Sustainable Organic Chemistry and Technology,Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium

[*]Corresponding author:
Technologiepark 914,
9052 Ghent, Belgium
Kevin.VanGeem@UGent.be

Chemical processes can benefit tremendously from fast and accurate effluent composition prediction for plant design, -control and -optimization. The Industry4.0 revolution claims that by introducing machine learning into these fields, both substantial economic and environmental gains could be achieved. The bottleneck for high-frequency optimization and process control is often the time necessary to perform the required detailed analyses of, e.g. feed and product. To resolve these issues, a framework of four deep learning artificial neural networks (DL ANNs) has been developed for the largest chemicals production process, i.e. steam cracking. The proposed methodology allows both a detailed characterization of a naphtha feedstock and a detailed composition of the steam cracker effluent to be determined, based on a limited number of commercial naphtha indices and rapidly accessible process characteristics. The detailed characterization of a naphtha is predicted from three points on the boiling curve and the PIONA characterization. If unavailable, the boiling points are also estimated. Even with estimated boiling points the developed DL ANN outperforms several established methods such as maximization of Shannon entropy and traditional ANNs. For feedstock reconstruction, a mean absolute error (MAE) of 0.3 wt% is achieved on the test set, while the MAE of the effluent prediction is 0.1 wt%. When combining all networks – using the output of the previous as input to the next – the effluent MAE increases to 0.19 wt%. Besides the high accuracy of the networks, another major benefit is the negligible computational cost required to obtain the predictions. On a standard Intel i7 processor, predictions are made in the order of milliseconds. Commercial software such as COILSIM1D performs slightly better in terms of accuracy, but the required CPU time per reaction is in the order of seconds. This tremendous speed-up and minimal accuracy loss, makes the presented framework highly suitable for continuous monitoring of difficult-to-access process parameters and the envisioned, high-frequency RTO strategy or process control. Nevertheless, the lack of fundamental basis implies that fundamental understanding is almost completely lost, which is not always well-accepted by the engineering community. Additionally, performance of the developed networks drops significantly for naphthas that are highly dissimilar to those in the training set.

Keywords: *Artificial Intelligence, Deep Learning, Steam Cracking, Artificial Neural Networks*

1. INTRODUCTION

With the majority of light olefins being produced via steam cracking – both today and in the foreseeable future [1] – it is important to take advantage of new technological developments and innovations. One such development that has taken the world by storm in the past few years is artificial intelligence (AI). AI has been widely adopted in several fields

such as strategic gaming [2, 3], natural language processing [4, 5] and autonomous cars [6, 7]. More recently, AI techniques have found their way into chemical (engineering) research [8]. Slowly, but steadily, AI is also making its way into industrial manufacturing and production processes [9]. Admittedly, the bulk chemical industry has been relatively conservative in this transition compared to *e.g.* the automotive sector. The upcoming technological revolution has been termed Industry 4.0, and is expected to redefine the limits of production [10-14]. Examples of the use of AI in chemistry include, among others, drug discovery [15, 16] and -synthesis [17, 18], and computational chemistry [19]. As indicated by the examples above, AI techniques excel at tackling highly complex and non-linear problems. Therefore, application of these methods to the modeling of the reactor section of the steam cracking process, being complex and non-linear itself, will deliver models which are expected to outperform traditional detailed kinetic models both in execution speed and accuracy. With increasing complexity and performance of real-time-optimization (RTO) systems – both in steam cracking and other industries [20-22] – the necessity for detailed inputs increases as well. While technically feasible, the use of comprehensive, on-line two-dimensional gas chromatography (2D-GC or GC×GC) for detailed stream characterization has not found its way into industry [23], due to its labor-intensive and time-consuming data processing. Hence the detailed compositions required in RTO systems are usually obtained via sampling and off-line analyses. These time-consuming analyses result in RTO systems that perform only one optimization step every few hours [24]. The above does not imply that on-line characterization techniques are not applied in industry, but the employed techniques for on-line characterization often relay much less detailed information than comprehensive GC×GC. Besides their value to RTO, detailed knowledge of reactor in- and output compositions is crucial to safe and efficient operation. Additionally, the development of accurate reactor models relies heavily on the level of detail of the feedstock and effluent characterization. The above implies the necessity for both feedstock reconstruction and reactor modeling algorithms. There is no lack of work on either of the topics, but only few approaches incorporate AI. Hudebine et al. [25, 26] and later Van Geem et al. [27] used entropy maximization methods with great success in feedstock reconstruction of various petroleum fractions. In reactor modelling the use of increasingly detailed kinetic models dominates other methods due to their capability to extrapolate beyond the ranges of predefined training sets [28-35]. One frequently used AI tool are artificial neural networks (ANNs) [36]. This form of biomimicry is a simplified mathematical representation of the neural network of the human brain as illustrated in Fig. 1 [37].

An example of the use of AI on the side of feedstock reconstruction is the work by Pyl et al. They developed an ANN to determine the detailed, molecular composition of naphthas typically used in cracking processes, based on their PIONA composition and boiling point curve [38]. Niaei et al. [39] and later Sedighi et al. [40] used ANNs to model reactor effluent compositions, but did so for a given feedstock. Ghadrdan et al. tackled this feedstock hiatus in a qualitative way by introducing a set of 9 feed type parameters to the ANN model [41]. While indisputably powerful tools, traditional ANNs and more classical machine learning techniques rely on the developer identifying the correct features that describe the problem. In this work, a deep learning (DL) approach will be applied to the problems of feedstock reconstruction and reactor effluent prediction for naphtha feedstocks.

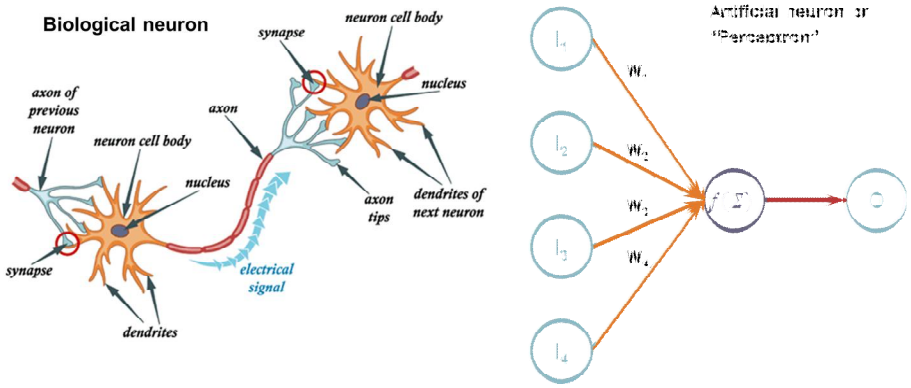

Fig. 1. Analogy between a biological neuron (a) and an artificial neuron or perceptron (b), after Jahnavi [37]

Deep learning further exploits the power of ANNs by relying on the network itself to identify, extract and combine the inputs into abstract features which contain much more pertinent information to solving the problem, *i.e.* predicting the output, as illustrated in Fig. 2 [42, 43]. The idea is that this additional level of abstraction improves the capability of the network to generalize to unseen data and hence outperform traditional ANNs on data outside of the network training set.

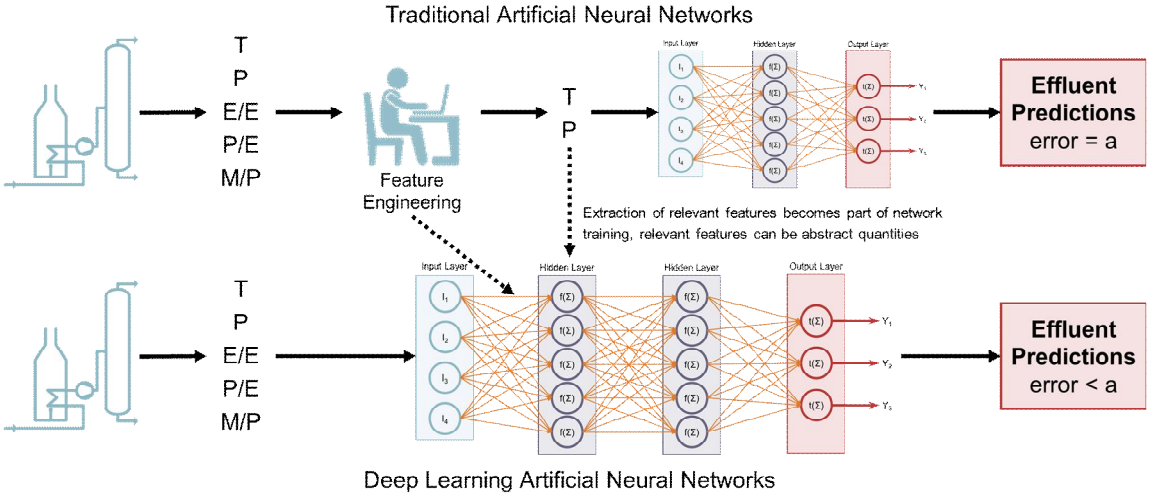

Fig. 2. Shallow artificial neural network compared to a deep learning ANN, after Seif [43]

In what follows four interacting DL ANNs are described, with as final goal achieving predictive accuracy on the steam cracker reactor effluent composition, using a limited number of commercial indices of the feedstock as input.

Fig. 3 illustrates this interacting DL ANN framework. 'Network 1' uses the most basic inputs – PIONA, density and vapor pressure – as input to predict the initial-, mid- and final boiling points. 'Network 2' uses these predicted boiling points, in combination with the previously specified PIONA to make a detailed reconstruction of the feedstock, which then can be used as input to 'Network 3'. This network predicts a detailed composition of the effluent. 'Network 4' serves as an extension and check for networks 1 and 2. Using a detailed PIONA characterization of a naphtha, it predicts its density, vapor pressure and the three aforementioned boiling points.

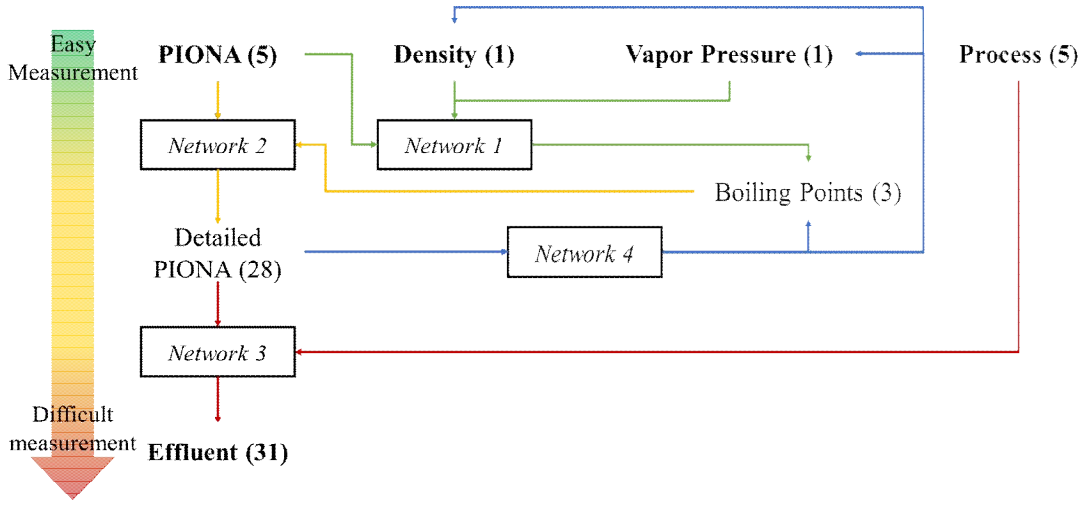

Fig. 3. Schematic overview of the interaction of the different variables and the four networks in the DL ANN framework

Before presenting the architecture of the individual DL ANNs in section 0, the theory of ANNs is briefly discussed and some comments concerning the data are given in section 0 and in the supporting information. In section 0, the results of the trained networks are discussed and compared to other reconstruction and prediction methods including support vector machine regression (SVR) and random forest regression (RF). In the final section we give a brief summary and comment on future prospects of this promising approach for steam cracking effluent prediction.

2. METHODS AND DATA

2.1. Deep Learning Artificial Neural Networks

The mathematical aspects of (DL) ANNs are similar, therefore in this section no distinction will be made between traditional ANNs and DL ANNs [44]. The relationship between the input '***i***' and output 'o' of a single perceptron is given by eq.1. All inputs are weighted by their respective weights 'w_j' and then summed. A constant bias term 'b' is added to this weighted sum. The activation function f introduces non-linearity into the network. Commonly used activation functions are sigmoid, hyperbolic tangent, rectified linear unit (ReLU) and softmax functions. More information on these activation functions can be found in section 1.1 of the supporting information. The equation for a single perceptron is easily extended to eq.2 to describe a full layer of the network, where ***W*** is the weight matrix of the layer. Each perceptron can have its own bias parameter. The entire network is finally described mathematically by repeatedly applying eq.2, which yields eq.3 for an ANN with one input layer, one hidden layer and one output layer, '***y***'.

$$o = f\left(\sum_j w_j \cdot i_j + b\right) = f(\boldsymbol{w} \cdot \boldsymbol{i} + b) \qquad \textbf{eq.1}$$

$$\boldsymbol{o} = f(\boldsymbol{W} \cdot \boldsymbol{i} + \boldsymbol{b}) \qquad \textbf{eq.2}$$

$$\boldsymbol{y} = f_2(\boldsymbol{W}_2 \cdot f_1(\boldsymbol{W}_1 \cdot \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) \qquad \textbf{eq.3}$$

The ANNs in this work are trained via back-propagation algorithms [44, 45], which update the network layer weights by passing down the error from one layer to the next, starting at the output. A gradient descent optimization approach is used to minimize a certain objective function. Frequently used error metrics in the objective function are (root) mean squared deviation (RMSD), mean absolute error (MAE) and mean absolute percentage error (MAPE). Typically several iterations through the complete training set are required to optimize the weights. One such iteration is termed an epoch. Within one epoch the training set is further split into several batches. The network weights are updated once per batch. A small batch size, *i.e.* a limited number of samples per optimization step, results in faster training in terms of the number of required epochs, but slower training in terms of computing time per epoch. Moreover, a smaller batch size results in poorer gradient estimates, reducing the stability of the optimization.

In ANNs, a distinction can be made between overfitting and overtraining of the network [46]. Overfitting occurs when the network becomes too complex, *i.e.* too many layers or too many nodes per layer are used. According to the universal approximation theorem, for any function an ANN can be found that approximates the data with any desired accuracy [47]. Overtraining on the other hand pertains to the number of training epochs. If the training data is shown to the network too often, it will start 'memorizing' the data, *i.e.* it will attempt to predict the exact output values, rather than the ones expected from the generalized trend in the data. This is illustrated by a simple example. Assume two variables are linearly related. In the dataset, one data point does not follow this linear trend, e.g. due to a measurement error. After a few training epochs, the network will have recognized the linear trend. The sum of squares however is still high due to the off-trend data point. During training, the sum of squares is minimized. Consequentially, in each following epoch, the network will start describing a trend that is increasingly less linear as after seeing the off-trend data point multiple times, it "believes" that that point is on-trend too. Overtraining can be ascertained by monitoring the objective function or network accuracy of both training and validation data set. While for the training set the objective function will typically follow a decreasing trend with increasing number of epochs, the objective function for the validation data will start to increase again at some point. From this point onward, the network is being overtrained. The above issues can be remedied, a.o. by using dropout during training [48, 49]. In this technique, during each batch of data, a randomly selected fraction of the network nodes is temporarily eliminated from the network. In this way, each neuron must individually learn characteristics – it cannot rely on neighboring neurons to capture information. All networks in this work use a dropout ratio of 0.5. The trade-off for the reduced overfitting with dropout, is that the network learns more slowly as only half the weights are updated in each step. Other regularization techniques such as L1- and L2 regularization [50] of the objective function have not been evaluated in this work as the constructed networks perform well on the test data.

The python deep learning library Keras [51] with Tensorflow backend [52] and GPU acceleration is used to train the artificial neural networks.

2.2. Data Analysis

2.2.1. Naphthas

From the work of Pyl et al. [38], a set of 272 detailed, industrial naphtha compositions is available. Available naphtha properties are density, vapor pressure, three ASTM D86 boiling points: initial boiling point (IBP), mid boiling point (BP50) and final

boiling point (FBP) and detailed paraffin, isoparaffin, olefin, naphthene and aromatic (PIONA) fractions per carbon number. Figure S-16 shows the correlation matrix of the available data. It is observed that vapor pressure and initial boiling point are strongly correlated, as are the density and mid boiling point. The end boiling point is less strongly correlated to density and vapor pressure, but significant correlation to the mid boiling point is present. This correlation will influence the architecture of the network to predict the boiling points from the vapor pressure and density of the naphtha, which will be discussed in paragraph 0.

Along the same lines of the work of Pyl et al. [38], a principal component analysis (PCA, details in S-1.2 [53]) is performed on the ten input variables of the dataset: IBP, BP50, FBP, density, vapor pressure and PIONA.

Fig. 4 summarizes the PCA results. From

Fig. 4 (a) it is concluded that the (training) dataset is described by three components. The scores of the inputs on the first two of these principal components (PC), shown in

Fig. 4 (b), confirm the findings from the correlation analysis. The high correlation observed between *e.g.* density and BP50 translates into parallel vectors in the PC-space. Though having opposite directions, the IPB and vapor pressure present similar behavior.

A second analysis based on PCA is performed on the test set. As ANNs rely only on the training and validation datasets during training, it can be expected that only test data which resembles the training and validation data, will yield accurate results. One measure to determine the resemblance of a data point to the training set is the Mahalanobis distance (MD) [54, 55]. In the PC space the MD is calculated via eq.4.

$$MD^2 = \mathbf{z}^T \cdot \mathbf{\Lambda'}^{-1} \cdot \mathbf{z} \qquad \textbf{eq.4}$$

z is the representation of the input in the PC space and contains the scores of the original input on each of the PCs. **Λ** is the matrix of all eigenvalues, *in casu* a 10×10 diagonal matrix. **Λ'** is the reduced 3×3 eigenvalue matrix and contains only the eigenvalues corresponding to the three selected PCs. Naphthas with a high MD can be considered outliers and can hence be expected to result in poorer predictions.

Fig. 4 (c) and (d) indicate the test set distribution in the PC space. The dotted line corresponds to a Mahalanobis distance of 2.5 and represents a probability of 90 % that a naphtha situated within the ellipsoid is within the range of the training set. This value of 2.5 for the Mahalanobis distance is used as critical distance to consider whether the corresponding naphtha is an outlier or not. One naphtha is observed to have a Mahalanobis distance of 5.08 and is indicated in red on

Fig. 4. In conclusion, this analysis indicates that predictions should be good in general, but may be off for the aforementioned naphtha.

2.2.2. Effluent Composition

Access to detailed, industrial steam cracker effluent compositions is highly restricted. Therefore, the state-of-the-art reactor simulation software tool COILSIM1D by Van Geem et al. [30, 56, 57] is used to obtain the required effluent characterizations. COILSIM1D has been validated against large amounts of proprietary data and is used in industry for detailed steam cracker simulations, indicating that it is a reliable and accurate tool. Hence, the obtained results are trusted to be an adequate replacement of the unavailable experimental or

industrial data. This approach of using simulation data as replacement for unavailable and/or limited experimental data has become common practice in other fields, especially in the prediction of thermodynamic properties of molecules and reaction kinetics [58-63]. The use of simulated data as training data, the difficulty in obtaining experimental data and the necessity of accurate input and output data underlines both the continuing importance of detailed, fundamental models for the simulation and understanding of these processes and the critical necessity for high-accuracy experimental techniques.

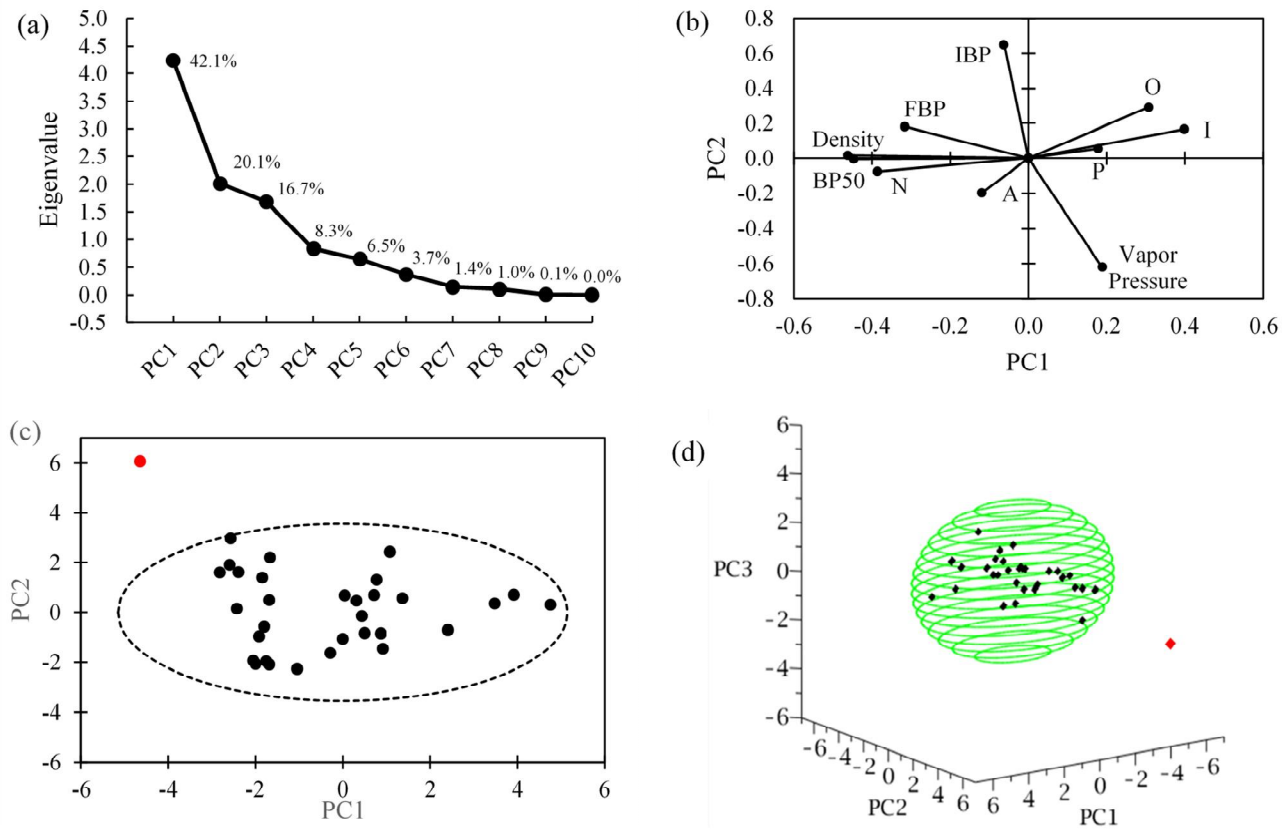

Fig. 4.
a) Eigenvalues and explained variance by the principal components;
b) Decomposition of the inputs along the first and second principal component (score plot);
c), d) PC representation of the naphtha test set. Outliers indicated in red

COILSIM1D can predict up to 100s of individual chemicals in the output. The majority of these components however are of minor importance to the overall operation of a steam cracker. Therefore 28 (pseudo-)components are identified. These comprise several molecular components, such as ethylene, propylene, benzene, hydrogen and butadiene and lumped components such as C_7 isoparaffins and C_{10+} aromatics. The full list of components can be found in S-2. Two sets of simulations are run.

The first set comprises a total of 13600 simulations and will be used to train and test the network to predict detailed effluent compositions. A different naphtha composition is used for each simulation. The different naphtha compositions are obtained from the dataset described in paragraph 0, but random variations of 0-10% have been introduced in the concentrations. Each naphtha is combined with a set of different process conditions. These

process conditions are the coil outlet pressure (COP) coil outlet temperature (COT). Figure S-17 shows that both the naphtha compositions and process conditions cover a wide range of the variable space in a uniform way. A single reactor and furnace configuration is used for all simulations. It will be shown later that the exact reactor configuration is of minor importance. A PCA on the new dataset is performed to identify potentially problematic cases.

Fig. 5 (a) indicates that the dataset is described well by six principal components. When projecting the test dataset onto the first three dimensions of the PC space, a small amount of inputs are observed to be situated outside of the ellipse encompassing 90% of the training data and corresponding to a Mahalanobis distance of 3.3. Again, this indicates that good overall performance on the test set can be expected, with a limited number of poor predictions.

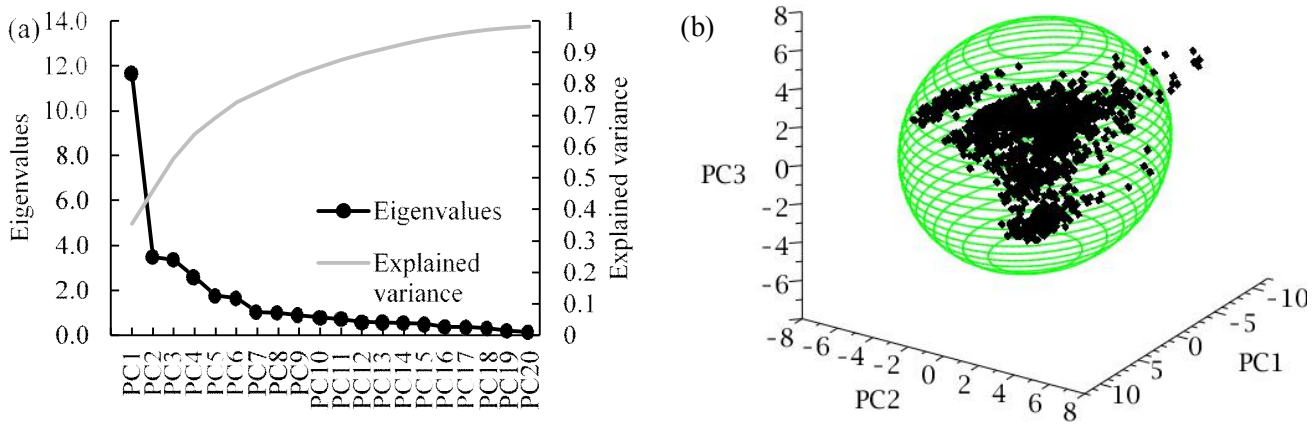

Fig. 5.
a) Eigenvalues and explained variance for the first 20 principal components in the PCA of the effluent dataset;
b) Effluent test data in the PC space reduced to three dimensions

The second set consists of 1587 additional simulations and is used to test the full workflow and combined performance of the networks. The same reactor and furnace configurations as for the previous simulations are used. 32 naphtha compositions are considered in this set, corresponding to the test set of networks 1,2 and 4, such that no training data is ever used during testing. Each of these naphthas is extended by a set of process conditions in fixed intervals. 10 COTs are considered in the range between 750°C and 950°C. Similarly 5 COPs between 1.7 and 2.3 bar are accounted for. Although this results in a more grid-like coverage of the variable space, it is sufficient for testing purposes.

3. SETUP OF THE ARTIFICIAL NEURAL NETWORKS

3.1. From Density and Vapor Pressure to Boiling Points

The aim of this work is to develop a set of algorithms that allow a user to obtain a detailed prediction of the steam cracking reactor effluent, using only readily available descriptors. As detailed predictions are more reliable when using detailed feedstock characterizations, a first step in the algorithm is reconstructing the feedstock from its commercial descriptors. Based on previous work by Van Geem and co-workers [27, 30-32, 38], it is apparent that at least some points on the naphtha boiling point curve are required to

successfully reconstruct the naphtha composition. However, boiling points are difficult to measure on-line as a single ASTM 86 compliant measurement can take 30-45 minutes [64]. Consequentially, they are not considered as "readily available". This makes estimating three important points on the boiling point curve, the first useful step towards predicting the effluent composition. The prediction of the boiling points is based on the density, vapor pressure and basic PIONA characterization of the naphtha. Based on the (cor)relations between the feed parameters described in paragraph 0, a network is constructed, of which the architecture is shown in Fig. 6. Due to the strong correlation of both the IBP and BP50 with the FBP , the vector containing the estimates for the IBP and BP50 is concatenated with the first hidden layer. This allows the network to use the predictions for the IBP and BP50 directly during the prediction of the FBP. The first hidden layer is chosen over the input layer as in the deep learning approach, the network is considered to learn the most relevant representation of the input towards predicting the output in this first hidden layer. Henceforth, this network will be referred to as "network 1". To increase the stability and performance of the network, all inputs and outputs have been normalized to the range of the dataset. The maxima and minima on which each variable has been normalized are listed in Table 1. The dataset of 272 naphthas is split into training-, validation- and test set according to a 80/8/12 split. The validation set is used to tune the hyperparameters of the network, *in casu* the number of nodes in the hidden layers, the batch size, the activation functions and the number of training epochs. In general "hyperparameters" denotes all parameters of the network except for the node weights and biases, which are referred to as the network parameters. The optimal combination is searched for heuristically. More detailed information on this search is given in S-3.1. The test set is used for evaluation of the final, optimized network.

The resulting hyperparameters are shown along with the architecture in Fig. 6. Additional figures comparing the performance of the network with different hyperparameters can be found in S-3.2. The mean absolute error (MAE) is preferred to the mean squared error as training objective function as, given the considered hyperparameter grid (cfr. S-3.1), the finally chosen network is observed to have the lower mean squared error. A detailed explanation for this specific network is given in S-3.2. Due to the normalization of the individual components, all outputs are of a similar order of magnitude. The use of the mean absolute percentage error (MAPE) is therefore not considered beneficial to the network accuracy. The best performance in terms of MAE is achieved with a batch size of 8 and after 1181 training epochs. The final network – using the optimized hyperparameters – is trained on both the training and validation data after which the network is validated against the unseen test data.

3.2. Feedstock Reconstruction

The second network in the framework uses the PIONA composition of the naphtha and the boiling points to reconstruct the detailed composition of the feedstock. For training the network, the experimental boiling points are used as input. In line with the work of Pyl et al. [38] 28 different pseudo-components are estimated, corresponding to the detailed PIONA matrix in Figure S-18. Additional distinction is made between xylenes and ethylbenzene and cyclohexane and methyl-cyclopentane in respectively the A_8 and N_6 categories. The inputs are normalized along the same procedure as in the previous paragraph, with the ranges listed in Table 1.

Table 1. Range for input and output variables of network 1

Variable [unit]	Minimum value	Maximum value
IBP [K]	303	328
BP50 [K]	323	398
FBP [K]	348	463
Density [-]	0.65	0.75
Vapor pressure [kPa]	27.6	84.9
P [wt%]	27.5	50
I [wt%]	25	52.5
O [wt%]	0	1
N [wt%]	5	35
A [wt%]	0	17

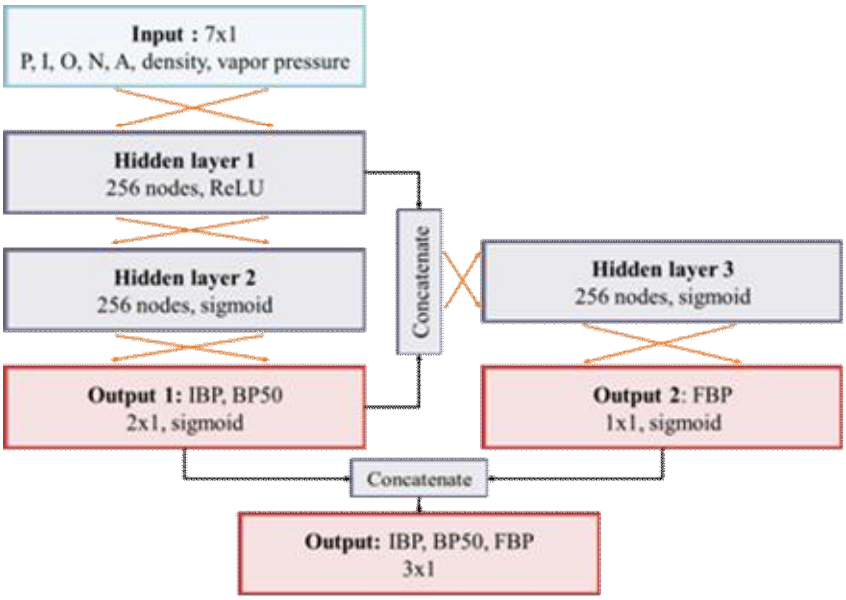

Fig. 6. Architecture of network 1, for predicting the IBP, BP50 and FBP, based on the PIONA composition, vapor pressure and density of the naphtha

For the outputs, a different normalization procedure is applied. The absolute concentrations of components in the different categories span a very wide range. The mass fraction C_5 and C_6 components can be as high as 35 wt%, while the olefin fraction can drop to 0.01 wt%. Attempting to directly predict all fractions at once, with a single softmax function, will result in a network that is difficult to train, especially considering the limited amount of available training data. The benefit of using a single softmax layer is that the outputs sum to one, corresponding to the physical nature of the desired mass fractions. Due to the wide range in mass fractions however, the five PIONA categories in the output are normalized individually according to the example for the paraffins in eq.5.

$$P_i^{Norm} = \frac{P_i}{\sum_j P_j} \qquad \textbf{eq.5}$$

By first splitting the output layer into five separate outputs, a softmax activation function can be used for each individual component category, with exception of the olefin mass fraction. Due to the fact that the total olefin concentration can be zero and the nature of the softmax activation function, the network is forced to incorrectly predict an olefin distribution that sums to one. This has a detrimental effect on the overall accuracy. Hence, for the olefin output layer, a sigmoid activation is used. The resulting multi-output architecture and optimized hyperparameters are shown in Fig. 7. In what follows, this network is referred to as "network 2". Again a train/validation/test split of 80/8/12 is used on the data. Section S-3.3 of the supporting information provides additional details on the optimization. In short, the MAE is chosen as network objective function. Due to the normalization per component class, the outputs do not span several orders of magnitude and hence do not require a relative cost function. For this network optimum performance is attained using a batch size of 16 and 45285 training epochs.

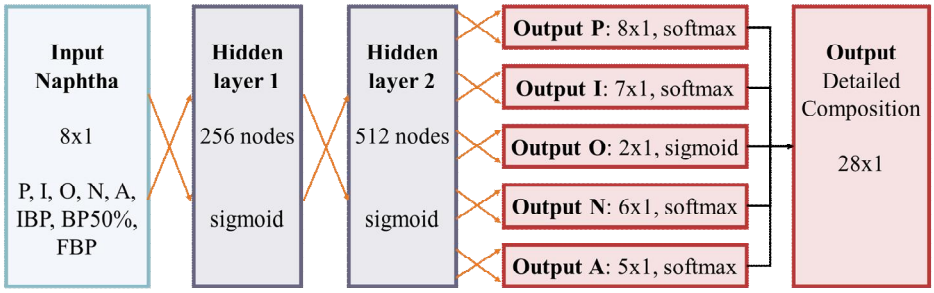

Fig. 7. Architecture of network 2 for reconstructing a more detailed feedstock composition starting from the PIONA characterization and boiling points

3.3. Detailed Effluent Prediction

The third network takes a detailed PIONA composition (28 pseudo-components) and five process characteristics as input to predict a detailed molecular composition of the steam cracker reactor effluent. As mentioned in paragraph 0, an adapted dataset is used for this network, containing 50 times more data points than the set used for the previous networks. The components considered in the detailed PIONA composition are the same as in Figure S-18. From previous work by Van Geem et al. [31], five process descriptors are identified. The first two – COT and COP – have already been used for the generation of the dataset. The remaining three are the product ratios of ethylene to ethane (E/E), propylene to ethylene (P/E) and methane to propylene (M/P). In their work, it is proven that for a given naphtha, the effluent composition is fully defined by just two of these descriptors. However, Figure S-19 learns that a more accurate model is obtained when including all five descriptors in the input. Three contributions to this increase in accuracy can be identified. For one, by using the aforementioned product ratios as input, the model must predict three fewer outputs, as methane, ethane and propylene mass fractions can be calculated from the prediction of the ethylene mass fractions. Secondly, by including multiple descriptors that essentially describe the same process parameters of temperature and pressure, the model becomes robust to errors in the input as the uncertainty is spread over multiple inputs. The most important reason can be traced back to the power of deep learning networks illustrated in

Fig. 2. By training the multi-layer network on multiple inputs, it is given the freedom to extract that information from the inputs which it finds to be most pertinent to solving the presented problem of predicting the effluent composition. Training the model using only *e.g.* COT and COP, does not make full use of the potential of deep learning. By manually selecting or engineering the network inputs and eliminating certain process descriptors from the network input, potentially useful information in the data is never shown to the network. In conclusion, all five identified descriptors are included in the network input.

Table 2. Range for process-related input variables of network 3

Variable [unit]	Minimum value	Maximum value
COT [K]	948	1318
COP [bara]	1.36	2.74
Ethylene/Ethane ratio [-]	2	37
Propylene/Ethylene ratio [-]	0	1.4
Methane/Propylene ratio [-]	0	35

The values for COT, COP, E/E, P/E and M/P are normalized on the ranges given by Table 2. Due to a mismatch in size between the inputs, the first layer is split into a process and a feedstock feature layer, yielding a more advanced DL ANN than the regular densely connected ones. This split allows for the extraction of independent, equally long, relevant feature vectors for both inputs. As it is not the complete effluent spectrum that is predicted by the network, the sum of the outputs should not equal one. Hence a softmax activation function cannot be applied in the output layer and a sigmoid activation is utilized instead, taking into account that the component fractions are bounded by zero and one. The final architecture and hyperparameters can be seen in Fig. 8. In this case, the mean absolute percentage error is chosen as objective function. Justification for this choice is given in section S-3.4 of the supporting information. The network is further referenced in this work as "network 3". For this dataset, a train/validation/test split of 81/9/10 is applied. The network reaches the best performance using a batch size of 8 and 2744 training epochs. Additional information on the optimization of the hyperparameters is given in S-3.4.

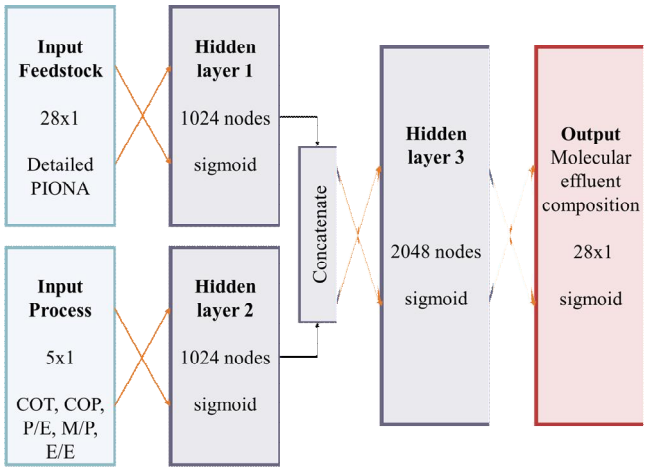

Fig. 8. Architecture of network 3 used to predict the molecular effluent composition based on the detailed feedstock composition and five process descriptors

3.4. Property Estimation

A final network in the framework serves as a check for the first two. Based on a detailed naphtha composition, the density, vapor pressure, IBP, BP50 and FBP are estimated. The dataset is identical to the one used for the reverse operation by networks 1 and 2. Given an accurate reconstruction, the predicted properties of a reconstructed naphtha should not differ much from those reported for the true naphtha. One could argue that the best results are obtained by simultaneously optimizing the four networks. However, given the limited size of the dataset, training such a complex network with multiple feedback loops is considered unfeasible at worst and inaccurate and non-generalizing at best. The fourth network – "network 4" – has a straightforward, two-layer architecture, with 28 inputs and 5 outputs, illustrated in Fig. 9 along with the optimized parameters. For similar reasons as for networks 1 and 2, the MAE is chosen as loss function. The 28 inputs are the same components accounted for in the reconstruction algorithm and are listed in Figure S-18. The sum of the 28 inputs is normalized to one, whereas the outputs are normalized according to the same ranges as previously mentioned in Table 1. A batch size of 8 and 5385 training epochs are found to yield the best performance. Additional information on the optimization is provided in S-3.5.

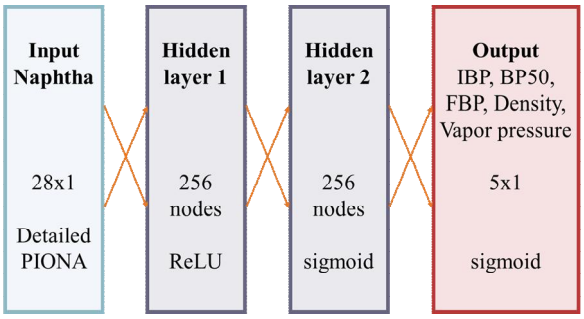

Fig. 9. Architecture of network 4 to predict naphtha properties from a detailed PIONA characterization

4. RESULTS AND DISCUSSION

4.1. Feedstock

The performance of the network to predict the initial-, mid- and final boiling points is shown in Fig. 10. Overall, the network performs very well, with only two notably poorer predictions, each for a different naphtha. They are indicated in red and green on the figure. The calculated Mahalanobis distance for the predictions in red is 1.82, which is below the critical value of 2.5 (cfr. 0), so accurate predictions are expected. The cause of this high error is discussed further on. The naphtha to which the green predictions correspond, is situated at a Mahalanobis distance of 5.08, which corresponds to a probability of 2.10^{-5} that the hypothesis of it belonging to the training set holds, for an F-statistic with (3, 237) degrees of freedom. The poorly predicted IBP is outside of the normalization range of the output, *cfr*. Table 1, indicating that the network has to predict a value greater than one, which is impossible by the construction of the network. For the other two boiling points however, the model makes very accurate predictions, despite the strong dissimilarity of the naphtha with the training dataset. Three other naphthas have a Mahalanobis distance greater

than the threshold value of 2.5. The predictions for these naphthas deviate up to 10 K from the experimental value. This shows one of the pitfalls of deep learning or any other type of regression; inputs that are very dissimilar to the those in the training set, will likely result in poorer predictions.

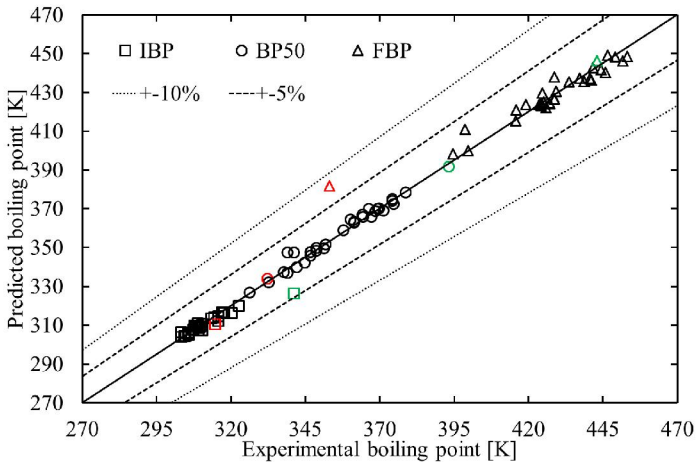

Fig. 10. Parity plot for network 1 – prediction of the IBP, BP50 and FBP from PIONA, density and vapor pressure. The points indicated in green are predictions for a naphtha with a Mahalanobis distance of 5.08, those in red for one of 1.82

Table 3. Statistical metrics of the performance of network 1 on the test set compared to work by Van Geem et al. [27]

	MAE [K]		RMSD		MAPE		Max deviation [K]	
	DL ANN	MSE	DL ANN	MSE	DL ANN	MSE	DL ANN	MSE
IBP	1.66	9.31	3.13	9.89	0.5%	3.0%	14.88	14.91
BP50	1.79	4.10	2.56	4.64	0.5%	1.2%	8.82	9.81
FBP	3.87	8.19	6.43	10.08	0.9%	1.9%	28.47	23.64

Table 3 shows that the predicted values deviate around 1% or 3 K from the experimental value on average, for all boiling points. This further substantiates the claim that it was not necessary to consider training the network on the MAPE. The accuracy of the network does not quite match experimental methods such as one with a maximum mean absolute error of 2.2 ± 1.4 K, reported by Ferris and Rothamer [65]. However, the deep learning ANN does perform better than the maximization of the Shannon entropy (MSE) approach used by Van Geem et al. [27]. This observation is not unexpected. The majority of the used test set, while never seen by the network during training, is situated within the ellipsoid corresponding to a Mahalanobis distance of 2.5 or a probability level of 0.9. Therefore, good performance of the network is expected even on the test set. Even for the data points situated outside of this critical ellipsoid, the DL ANN model still performs similarly to the MSE approach. This is supported by their similar maximal deviations.

A very high throughput can be achieved with the network – prediction of the boiling points of the 32 test naphthas took 137 ms on a 2.7 GHz Intel i7-6820HQ CPU, or just over 4 ms per naphtha. Unfortunately, an equivalent speed test with the method of Van Geem et al. was not possible, as the estimate of the boiling points is reported as part of the feedstock reconstruction, though given the combined time of 25s, the DL ANN can safely be assumed to be faster.

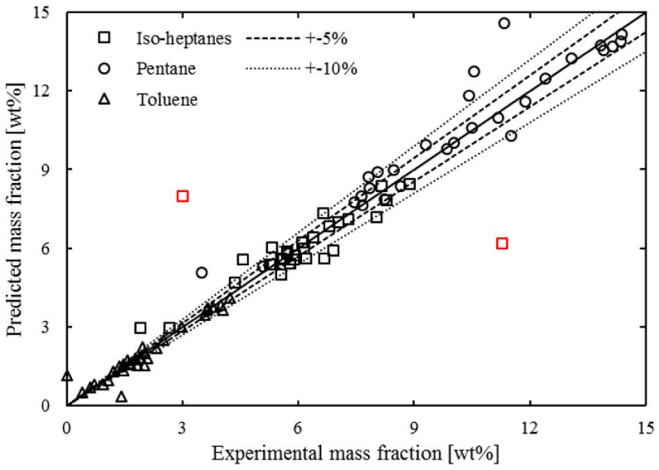

Fig. 11. Network 2 performance on selected components of the output

Fig. 11 shows the performance of network 2 on a selected number of components in the output. Parity plots for all components in the output can be found in Figure S-13. In general the performance is good over the entire range of concentrations. The network achieves an overall mean absolute error of 0.31 wt%. Two outlying predictions are singled out in red. In paragraph 0, a lack of correlation for the I_7 components with any of the other variables was mentioned. When leaving out the naphthas corresponding to the highlighted points, the correlation of the I_7 component group to other variables is found to increase by over 1%. As the left out data accounts for about 0.7% of the data, it can be concluded that they have a significant impact on the lack of correlation. The calculated MD for the naphthas is 2.27 ("naphtha A") and 1.82 ("naphtha B") respectively. Therefore there is no indication that the naphtha compositions are outside the scope of the training set. The above suggests that it is possible that a measurement error is causing the poor prediction. This possibility is further supported by the fact that nearly all off-trend predictions noticed for other components (*e.g.* P_4 and P_7) are the result of the same two problematic naphthas. A measurement error for one or more components could also help explain the poor prediction of the final boiling point of the naphtha highlighted in red in Fig. 10, as it is the same naphtha as "naphtha B". This highlights the critical importance of high-quality input, both for accurately training the network and for obtaining accurate predictions.

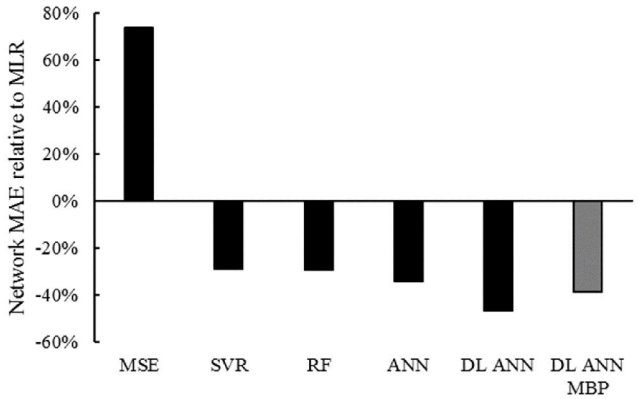

Fig. 12. Network mean absolute error (MAE) relative to that of the MLR model. DL ANN MBP: uses the modeled boiling points as input – combined performance of networks 1 and 2

The performance of network 2 is compared to previous work on feedstock reconstruction by K. Van Geem [27] and S. Pyl [38] and two additionally constructed models, in which the reconstruction algorithms are based on the following methods: MSE (Van Geem), multiple linear regression (MLR, Pyl), traditional ANNs (Pyl), support vector regression (SVR), and random forest regression (RF). The MLR approach – being the traditional method – is used as performance baseline. Table 4 shows the performance of the different models on the individual components of the output. Machine learning techniques such as SVR and (DL) ANNs show a significant improvement compared to the more traditional methods such as MLR and MSE. Fig. 12 shows the relative model performance in terms of MAE. The deep learning approach clearly outperforms all other models – network 2 attains an MAE that is just over half the MLR MAE and still 20% lower than the ANN MAE . Even when using the predicted boiling points based on the density and vapor pressure – combining network 1 and 2 – the DL ANN still performs noticeably better than all other tested models. While the MSE approach has a significantly higher MAE, its advantage is that it relies on a case-by-case optimization, *i.e.* the applicability of the method is less restricted to the range of a certain training set. In terms of required CPU time, the MSE method takes about 25 s to simulate both boiling points and reconstruct the detailed composition for the test set. Using network 1 and 2, the combined process only requires about one tenth of that time – 234 ms – on the same Intel i7 processor mentioned in the previous paragraph.

Network 4 also pertains to the feedstock as it estimates properties based on a known, detailed composition. The performance of this network is illustrated by the parity plots in Fig. 13. The singled-out predictions in Fig. 13 (a) correspond to those for "naphtha B" mentioned before. Again, the poor prediction for the vapor pressure could be the result of measurement errors during the compositional analysis of the naphtha. Table 5 shows the statistics of the network performance. The performance of the combination of networks 1,2 and 4 is also displayed in the table. There is a clear decrease in performance of the network when starting from the most basic commercial indices, however, reasonably accurate results are still obtained and the general trend of the properties is still predicted well.

Table 4. Mean absolute error [wt%] of different algorithms for the detailed reconstruction of naphthas, based on PIONA and boiling points

	MSE	MLR	SVR	RF	ANN	DL ANN	DL ANN MBP
P_4	1.75	0.52	0.44	0.60	0.50	0.52	0.47
P_5	2.28	1.16	1.03	1.17	0.97	0.65	0.58
P_6	1.16	1.10	0.95	0.80	0.71	0.71	0.95
P_7	1.15	0.63	0.48	0.50	0.47	0.47	0.60
P_8	0.66	0.50	0.39	0.31	0.29	0.25	0.33
P_9	0.57	0.32	0.26	0.26	0.26	0.20	0.23
P_{10}	0.27	0.22	0.10	0.11	0.11	0.10	0.09
P_{10}	0.05	0.06	0.04	0.03	0.03	0.02	0.02
I_4	2.40	1.11	0.85	0.99	0.82	0.58	0.65
I_5	1.63	1.40	1.03	0.91	0.85	0.83	0.96
I_7	2.40	1.02	0.80	0.80	0.84	0.66	0.72
I_8	1.41	0.62	0.45	0.38	0.44	0.32	0.42

I_9	0.63	0.47	0.32	0.30	0.32	0.28	0.33
I_{10}	0.52	0.44	0.29	0.28	0.25	0.19	0.20
I_{11}	0.11	0.10	0.05	0.04	0.04	0.03	0.03
O_5	0.01	0.04	0.02	0.02	0.05	0.02	0.02
O_6	0.04	0.03	0.01	0.02	0.02	0.01	0.01
N_5	2.20	0.20	0.15	0.16	0.14	0.16	0.17
N_6-1	1.48	1.07	0.55	0.43	0.53	0.43	0.46
N_6-2	1.48	1.07	0.55	0.54	0.53	0.35	0.46
N_7	2.18	0.84	0.65	0.80	0.56	0.58	0.69
N_8	0.56	0.60	0.45	0.39	0.31	0.28	0.41
N_9	0.93	0.46	0.42	0.34	0.34	0.30	0.34
A_6	0.61	0.54	0.56	0.50	0.30	0.28	0.31
A_7	0.81	0.45	0.27	0.37	0.26	0.19	0.22
A_8-1	0.36	0.56	0.29	0.25	0.26	0.16	0.16
A_8-2	0.36	0.56	0.10	0.08	0.26	0.06	0.07
A_9	0.58	0.38	0.24	0.26	0.39	0.17	0.17
Average	1.02	0.59	0.42	0.42	0.39	0.31	0.36

Table 5. Statistics on the performance of network 4 on the test set and on the reconstruction of the test set based on the vapor pressure and density of the naphtha

	MAE		MAPE		RMSD		Max deviation	
	Original	Reconstr.	Original	Reconstr.	Original	Reconstr.	Original	Reconstr.
IBP [K]	1.87	4.24	0.6%	1.3%	3.49	6.40	16.44	27.6
BP50 [K]	1.82	11.8	0.5%	3.3%	2.65	13.2	8.70	22.9
FBP [K]	4.35	9.93	1.0%	2.4%	5.73	13.0	13.28	35.3
Density [-]	0.001	0.02	0.2%	2.7%	0.002	0.02	0.005	0.03
Vapor pressure [kpa]	2.28	11.45	3.8%	17.3%	3.94	13.80	18.09	26.41

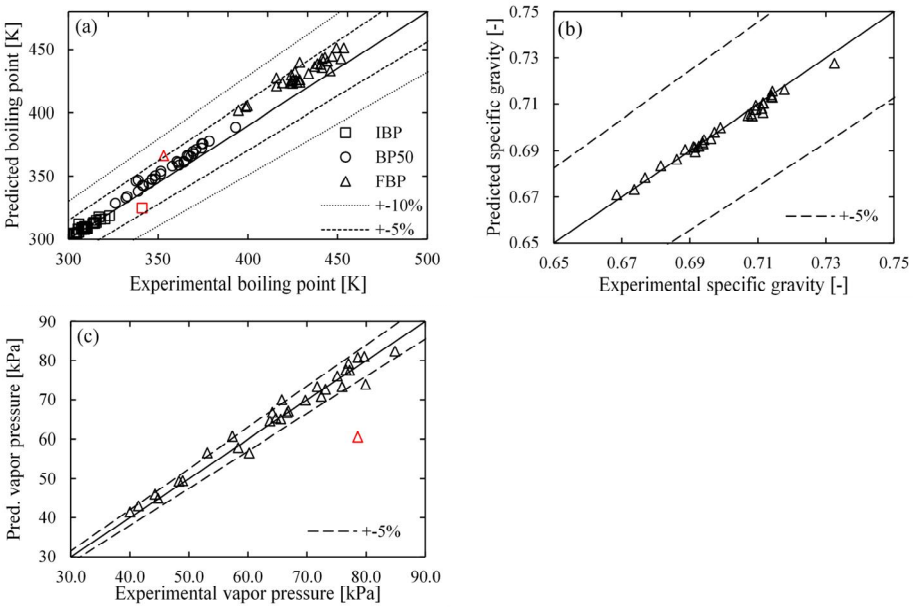

Fig. 13. Parity plots for the different outputs of network 4:
(a) IBP, BP50, FBP; (b) density as specific gravity; (c) vapor pressure.
The red data points correspond to "naphtha B"

4.2. Effluent

The performance of network 3 is first evaluated separately due to the use of a different training and test set. All following figures use a random selection of 10% of the 1360 data points in the test set to maintain the legibility of the figures. The statistical metrics are calculated on the full test set. Fig. 14 illustrates the network performance on four selected output components – ethene, butadiene, hydrogen and heavy aromatics. Parity plots for all other components can be found in Figure S-14, which shows that the network performance for two other major cracking products – methane and propene – is very similar to that for ethene shown in Fig. 14 (a). The network performs well on the entire range of mass fractions. For ethylene, butadiene and hydrogen, the mass fraction range is limited to about one order of magnitude. For the A_{10}+ pseudo-component, however, the mass fractions of the data set are spread out over nearly four orders of magnitude. By accurately predicting the mass fractions of the A_{10}+ pseudo-component across several orders of magnitude, the network demonstrates its predictive power. Table 6 shows the statistics for these four components specifically and the averages for all components. In general the network achieves an accuracy of 0.1 wt%, which is very high taking into account the minimal computational cost of the predictions. The entire test set of 1360 reactions is predicted in 1.716 s, or just 1.2 ms per prediction, once again on a standard Intel i7 laptop CPU. The state-of-the-art tool COILSIM1D requires several seconds to determine the detailed effluent composition for a single naphtha, indicating a tremendous speed-up for the DL ANN model. The (nearly) negligible computation times would allow such a network to be used in a larger RTO algorithm that is able to provide feedback to the process at a much higher frequency than current RTO algorithms. At this computation speed, even feed-forward process control applications are possible. The major benefit of this tremendous speed-up is, however, the ability to continuously monitor difficultly accessible process parameters with limited input. This facilitates anticipating sudden changes that might have a major (safety) impact on downstream operations.

In paragraph 0 it was mentioned that the exact reactor configuration is of secondary importance. Van Geem et al. [31] have proven that the composition of the reactor effluent for a given naphtha is defined by two severity indices accounting for outlet pressure and temperature, independently of the reactor geometry. Network 3 uses these severity indices – P/E and E/E – as input. Hence the performance of the network will be relatively independent of the reactor geometry and it can thus be used to obtain good predictions for any type of reactor. These findings are graphically supported by Figure S-21.

Table 6. Statistics on the performance of network 3 on selected components, on the test set

	MAE [wt%]	MAPE [%]	RMSD [-]
Ethylene	0.42	1.9	0.763
Butadiene	0.10	3.1	0.150
Hydrogen	0.02	1.8	0.029
A_{10}+ fraction	0.18	7.3	0.762
Average	0.13	7.3	0.416

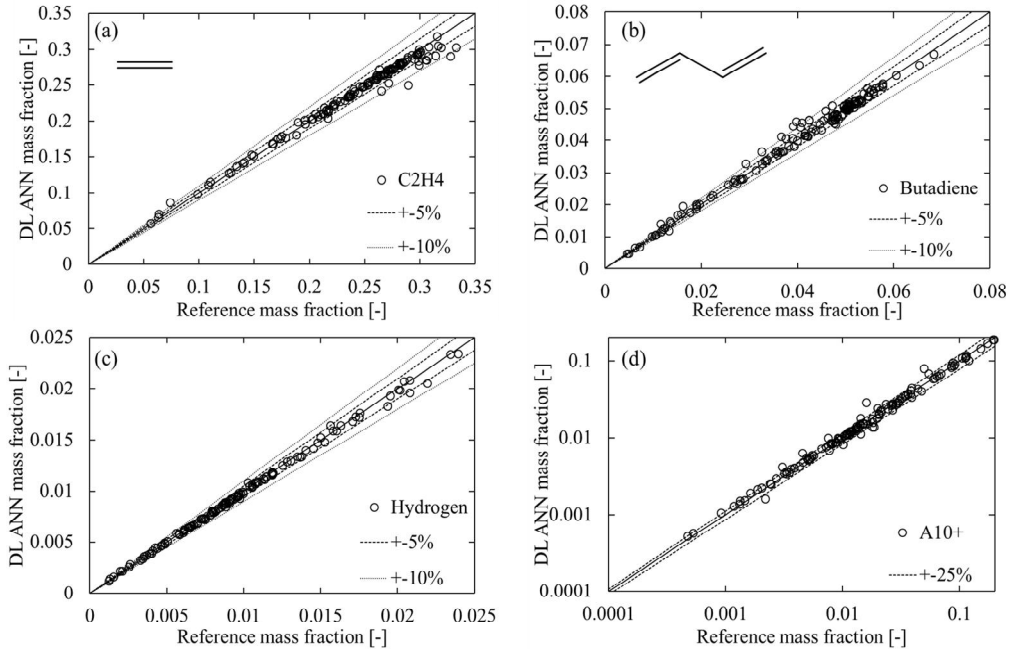

Fig. 14. Parity plots for the predictions by network 3 on four selected components: (a) ethylene; (b) 1,3-butadiene; (c) hydrogen and (d) C_{10}+ aromatics. 136 of the 1360 test set data points are displayed

4.3. Combined Effluent Prediction Performance

Finally the performance of the combination of feedstock reconstruction from easily and rapidly accessible indices and detailed effluent prediction is evaluated. This corresponds to evaluating the performance of the framework elucidated in

Fig. 3.

The computational cost to run the combined framework is still very low. The 1587 test cases are simulated in just under 3.25 s – 2 ms per reactor simulation, which is only a minimal increase compared to the time required to simulate the effluent from the detailed naphtha characterization. This indicates that the combined framework is at least computationally suited for integration in RTO algorithms, or even in direct process control.

Table 7. Statistics on the combined performance of networks 1,2 and 3, on selected components, on the test set

	MAE [wt%]	MAPE [%]	RMSD[-]
Ethylene	0.46	1.9%	0.594
Butadiene	0.16	3.9%	0.206
Hydrogen	0.02	3.2%	0.030
A_{10}+ fraction	0.95	35.1%	1.167
A_7+ fraction	0.43	8.9%	0.594
Average	0.19	15.0%	0.385

Upon comparing Fig. 14 to Figure S-20 and Table 6 to Table 7, a drop in performance for the combination of networks 1,2 and 3 is observed. For several components, such as ethylene, butadiene and hydrogen, the network accuracy is still very

high and close to the accuracy using the true naphtha composition. The network does have significant trouble correctly predicting the distribution between A_{7-9} and A_{10+}. The parity plot for the former can be found in Figure S-15, that of the latter in Figure S-20 (d). The concentration of the lighter aromatics is consistently overestimated while that of the heavier aromatics is consistently underestimated. When these two pseudo-components are further lumped into a single A_{7+} component, the network achieves an accuracy similar to the others, as shown in the next to last row of Table 7. A potential cause for this deviation could be a very slight, systematic underestimation of the aromatics at higher concentrations in the feedstock reconstruction. It is observed that a small variation in the aromatics content of the feedstock can significantly impact the formation of heavier aromatic compounds during the cracking process. This shows the importance of very accurate experimental data, as small measurement errors can significantly impact the results.

The clustering of the results in the parity plots of Figure S-20 and Figure S-15 is the result of the grid-like variation in the input. While the process conditions will influence the exact characteristics of the output, the naphtha composition is the main influence on the effluent composition. As only 32 different naphthas were considered for this dataset, it is not surprising that only certain regions of the effluent space are covered.

5. CONCLUSIONS AND OUTLOOK

A framework of four, interacting, deep learning artificial neural networks has been developed for the prediction of naphtha properties and detailed steam cracker effluent compositions, based on a limited number of commercial, or easily accessible naphtha characteristics and process descriptors. Each of the individual networks achieves excellent performance that rivals or outperforms the accuracy of typical on-line analysis equipment and commercially available tools such as COILSIM1D. Using two DL ANNs to reconstruct a detailed feedstock composition from the PIONA characterization of the naphtha and its density and vapor pressure, an average MAE of 0.36 wt% across 28 different (pseudo-)components is achieved. The effluent composition can be predicted with an average MAE of 0.13 wt% when using the true, detailed naphtha composition and an average MAE of 0.19 wt% when using a naphtha composition reconstructed from the above mentioned indices. This high, predictive accuracy, combined with very low computational costs – execution of the full framework takes place in the order of milliseconds – makes the developed networks very well suited for real-time monitoring of difficult-to-access process parameters. They are also suited for use in new RTO algorithms with a much higher frequency of process adjustments than current ones. At computational delays in the order of milliseconds, even application in feed-forward process control can be considered. While the presented networks have been trained on simulations for a specific configuration of the reactor and furnace, the inclusion of reactor-independent severity indices in the input makes the network itself reactor-independent. As a result the presented method is applicable to any type of reactor without loss of performance. The main disadvantage of DL ANNs is that the physical and interpretable meaning of the problem is lost. For detailed cause and effect analyses on the complex chemical mechanisms behind the process and process design, detailed kinetic models are still essential. The fact that the presented models have been trained on simulated data, further advocates the development of fundamental models. However, for many practical applications such as the above-mentioned RTO and process control, the combination of execution speed, accuracy, and ease of use are the main concerns. Due to the flexibility and predictive power of DL ANNs, several other aspects of

the steam cracking process that influence the plant optimization – e.g. coke formation – could be approached in a similar way in the future.

ACKNOWLEDGEMENTS

Pieter Plehiers acknowledges financial support from a doctoral fellowship from the Research Foundation - Flanders (FWO). Ismaël Amghizar acknowledges financial support from SABIC Geleen. The authors acknowledge funding from the COST Action CM1404 "Chemistry of smart energy and technologies. This work was funded by the EFRO Interreg V Flanders-Netherlands program under the IMPROVED project.

The authors also acknowledge the financial support from the Long Term Structural Methusalem Funding by the Flemish Government – grant number BOF09/01M00409.

COMPLIANCE WITH ETHICS GUIDELINES

All authors declare that they have no conflict of interest or financial conflicts to disclose.

NOMENCLATURE

Abbreviations

2D-GC	Two-Dimensional Gas Chromatography	
AI	Artificial Intelligence	
ANN	Artificial Neural Network (1 hidden layer)	
BP	Boiling Point	[K]
BP50	Mid Boiling Point	[K]
COP	Coil Outlet Pressure	[bar]
COT	Coil Outlet Temperature	[K]
CPD	Cyclopentadiene	
CPU	Central Processing Unit	
DL	Deep Learning (> 1 hidden layer)	
E/E	Ethylene/Ethane ratio	[-]
FPB	Final Boiling Point	[K]
GC×GC	Two-Dimensional Gas Chromatography	
GPU	Graphics Processing Unit	
IPB	Initial Boiling Point	[K]
M/P	Metane/Propylene ratio	[-]
MAE	Mean Absolute Error	
MAPE	Mean Absolute Percentage Error	
MBP	Modeled Boiling Point	[K]
MD	Mahalanobis Distance	
MLR	Multiple Linear Regression	
MSE	Maximization of the Shannon Entropy	
P/E	Propylene/Ethylene ratio	[-]
PC(A)	Principal Component (Analysis)	
PIONA	Paraffins, Isoparaffins, Olefins, Naphthenes, Aromatics	
ReLU	Rectified Linear Unit	
RF	Random Forest	
(R)MSD	(Root) Mean Square Deviation	
RTO	Real-Time-Optimization	
SVR	Support Vector Regression	

Greek and Roman Symbols

A	Matrix of eigenvectors
A_i	Aromatics with i carbon atoms
b	Perceptron/Layer bias
C_i	Hydrocarbons with i carbon atoms

d	(chosen) Dimensionality of the PC space
f	Activation function
$F_{\alpha, p, n}$	F-statistic with confidence level α, p degrees of freedom and n samples
i	Perceptron/layer input
i	Perceptron/layer input vector
I_i	Isoparaffins with i carbon atoms
n	Number of data points in dataset
N_i	Naphthenes with i carbon atoms
o	Layer output
o	Perceptron output
O_i	Olefins with i carbon atoms
P_i	Paraffins with i carbon atoms
S	Variance- covariance matrix of the dataset
w	Weight
W	Weight matrix for single layer
w	Weight vector for single perceptron
x	Model input
x	Model input vector
y	Model output
y	Model output vector
z	Input representation in the PC space
α	Probability level
Λ	Diagonal matrix of eigenvalues
λ	Eigenvalue
Λ'	Eigenvector matrix in the reduced-dimension PC space

REFERENCES

1. Amghizar, I.; Vandewalle, L. A.; Van Geem, K. M.; Marin, G. B., New Trends in Olefin Production. *Engineering* **2017,** 3, (2), 171-178.
2. Campbell, M.; Hoane, A. J.; Hsu, F.-h., Deep Blue. *Artificial Intelligence* **2002,** 134, (1), 57-83.
3. Gibney, E., Google AI algorithm masters ancient game of Go. *Nature News* **2016,** 529, (7587), 445.
4. Chowdhury, G. G., Natural language processing. *Annual Review of Information Science and Technology* **2003,** 37, (1), 51-89.
5. Yin, W.; Kann, K.; Yu, M.; Schütze, H., Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923* **2017**.
6. Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J., End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* **2016**.
7. Li, D.; Gao, H., A Hardware Platform Framework for an Intelligent Vehicle Based on a Driving Brain. *Engineering* **2018,** 4, (4), 464-470.
8. Maltarollo, V. G.; Honório, K. M.; Ferreira da Silva, A. B., Applications of Artificial Neural Networks in Chemical Problems. In *Artificial Neural Networks - Architectures and Applications*, Suzuki, K., Ed. InTech: Rijeka, 2013.
9. Day, C.-P., Robotics in Industry—Their Role in Intelligent Manufacturing. *Engineering* **2018,** 4, (4), 440-445.
10. Brettel, M.; Friederichsen, N.; Keller, M.; Rosenberg, M., How virtualization, decentralization and network building change the manufacturing landscape: An Industry 4.0 Perspective. *International Journal of Mechanical, Industrial Science and Engineering* **2014,** 8, (1), 37-44.
11. Lasi, H.; Fettke, P.; Kemper, H.-G.; Feld, T.; Hoffmann, M., Industry 4.0. *Business & Information Systems Engineering* **2014,** 6, (4), 239-242.
12. Ray Y. Zhong, X. X., Eberhard Klotz, Stephen T. Newman, Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering* **2017,** 3, (5), 616-630.
13. Zhou, K.; Taigang, L.; Lifeng, Z., In *Industry 4.0: Towards future industrial opportunities and challenges*, 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015; pp 2147-2152.

14. Yuan, Z.; Qin, W.; Zhao, J., Smart Manufacturing for the Oil Refining and Petrochemical Industry. *Engineering* **2017,** 3, (2), 179-182.
15. Zhang, L.; Mao, H.; Liu, L.; Du, J.; Gani, R., A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. *Computers & Chemical Engineering* **2018,** 115, 295-308.
16. Bajorath, J., Computer-aided drug discovery. *F1000Research* **2015,** 4, F1000 Faculty Rev-630.
17. Peplow, M., Organic synthesis: The robo-chemist. *Nature* **2014,** 512, (7512), 20.
18. Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F., SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling* **2018,** 58, (2), 252-261.
19. Goh, G. B.; Hodas, N. O.; Vishnu, A., Deep learning for computational chemistry. *Journal of Computational Chemistry* **2017,** 38, (16), 1291-1307.
20. Sedghi, S.; Huang, B., Real-Time Assessment and Diagnosis of Process Operating Performance. *Engineering* **2017,** 3, (2), 214-219.
21. Bogle, I. D. L., A Perspective on Smart Process Manufacturing Research Challenges for Process Systems Engineers. *Engineering* **2017,** 3, (2), 161-165.
22. Castillo, P. A. C.; Castro, P. M.; Mahalec, V., Global Optimization of Nonlinear Blend-Scheduling Problems. *Engineering* **2017,** 3, (2), 188-201.
23. 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, (43), 6623-6633.
24. Van Geem, K.; Marin, G.; Muñoz Gandarillas, A.; Zhang, Y.; Du, W.; Qian, F., In *Plant Wide Optimization for High Value Added Products-a Steam Cracking Case Study*, 30th Ethylene producers' conference (30th EPC), 2018.
25. Hudebine, D.; Verstraete, J. J., Molecular reconstruction of LCO gasoils from overall petroleum analyses. *Chemical Engineering Science* **2004,** 59, (22), 4755-4763.
26. Verstraete, J. J.; Revellin, N.; Dulot, H.; Hudebine, D., Molecular reconstruction of vacuum gasoils. *Prepr. Pap.-Am. Chem. Soc., Div. Fuel Chem* **2004,** 49, (1), 20.
27. 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, (9), 1020-1034.
28. 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, (1), 99-139.
29. Sadrameli, S. M., Thermal/catalytic cracking of hydrocarbons for the production of olefins: A state-of-the-art review I: Thermal cracking review. *Fuel* **2015,** 140, 102-115.
30. 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, (1), 79-94.
31. Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B., Two Severity Indices for Scale-Up of Steam Cracking Coils. *Industrial & Engineering Chemistry Research* **2005,** 44, (10), 3402-3411.
32. Van Geem, K. M.; Žajdlík, R.; Reyniers, M.-F.; Marin, G. B., Dimensional analysis for scaling up and down steam cracking coils. *Chemical Engineering Journal* **2007,** 134, (1), 3-10.
33. Van Geem, K. M.; Zhou, Z.; Reyniers, M.-F.; Marin, G. B., In *Effect of operating conditions and feedstock composition on run lengths of steam cracking coils*, AIChE Spring Meeting: Ethylene producers conference, Tampa, FL, Tampa, FL, 2009.
34. Green Jr, W. H., Predictive Kinetics: A New Approach for the 21st Century. In *Advances in Chemical Engineering*, Guy, B. M., Ed. Academic Press: 2007; Vol. 32, pp 1-50.
35. Van de Vijver, R.; Vandewiele, N. M.; Bhoorasingh, P. L.; Slakman, B. L.; Seyedzadeh Khanshan, F.; Carstensen, H.-H.; Reyniers, M.-F.; Marin, G. B.; West, R. H.; Van Geem, K. M., Automatic Mechanism and Kinetic Model Generation for Gas- and Solution-Phase Processes: A Perspective on Best Practices, Recent Advances, and Future Challenges. *International Journal of Chemical Kinetics* **2015,** 47, (4), 199-231.
36. Hopfield, J. J., Artificial neural networks. *IEEE Circuits and Devices Magazine* **1988,** 4, (5), 3-10.
37. Jahnavi, Introduction to Neural Networks, Advantages and Applications. https://www.deeplearningtrack.com/single-post/2017/07/09/Introduction-to-NEURAL-NETWORKS-Advantages-and-Applications

38. 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, (12), 3174-3188.
39. Niaei, A.; Towfighi, J.; Khataee, A. R.; Rostamizadeh, K., The Use of ANN and the Mathematical Model for Prediction of the Main Product Yields in the Thermal Cracking of Naphtha. *Petroleum Science and Technology* **2007,** 25, (8), 967-982.
40. Sedighi, M.; Keyvanloo, K.; Towfighi, J., Modeling of Thermal Cracking of Heavy Liquid Hydrocarbon: Application of Kinetic Modeling, Artificial Neural Network, and Neuro-Fuzzy Models. *Industrial & Engineering Chemistry Research* **2011,** 50, (3), 1536-1547.
41. Ghadrdan, M.; Mehdizadeh, H.; Boozarjomehry, R. B.; Darian, J. T., On the Introduction of a Qualitative Variable to the Neural Network for Reactor Modeling: Feed Type. *Industrial & Engineering Chemistry Research* **2009,** 48, (8), 3820-3824.
42. Szegedy, C.; Toshev, A.; Erhan, D., In *Deep neural networks for object detection*, Advances in neural information processing systems, 2013; pp 2553-2561.
43. Seif, G., I'll tell you why Deep Learning is so popular in demand. https://medium.com/swlh/ill-tell-you-why-deep-learning-is-so-popular-and-in-demand-5aca72628780 (08/2018),
44. Shamsuddin, S. M.; Ibrahim, A. O.; Ramadhena, C., Weight Changes for Learning Mechanisms in Two-Term Back-Propagation Network. In *Artificial Neural Networks-Architectures and Applications*, Suzuki, K., Ed. InTechOpen: 2013.
45. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J., Learning representations by back-propagating errors. *Nature* **1986,** 323, 533.
46. Tetko, I. V.; Livingstone, D. J.; Luik, A. I., Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences* **1995,** 35, (5), 826-833.
47. Hornik, K.; Stinchcombe, M.; White, H., Multilayer feedforward networks are universal approximators. *Neural Networks* **1989,** 2, (5), 359-366.
48. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., In *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012; pp 1097-1105.
49. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **2014,** 15, (1), 1929-1958.
50. Ng, A. Y., Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the twenty-first international conference on Machine learning*, ACM: Banff, Alberta, Canada, 2004; p 78.
51. Chollet, F., *Keras*, https://keras.io, 2015.
52. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M., In *Tensorflow: a system for large-scale machine learning*, OSDI, 2016; pp 265-283.
53. Jolliffe, I. T., *Principal component analysis*. Springer: 2011.
54. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L., The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **2000,** 50, (1), 1-18.
55. Mahalanobis, P. C., In *On the generalized distance in statistics*, National Institute of Science of India: 1936.
56. Van Geem, K. M.; Reyniers, M.-F.; Marin, G., In *Taking optimal advantage of feedstock flexibility with COILSIM1D*, AIChE Spring Meeting: Ethylene producers conference, 2008.
57. Vervust, A.; Amghizar, I.; Munoz, A.; Van Geem, K.; Marin, G., In *Full furnace simulations and optimization with coilsim1d*, American Institute of Chemical Engineers Spring Meeting (AIChE Spring 2016), 2016.
58. Paraskevas, P. D.; Sabbe, M. K.; Reyniers, M.-F.; Marin, G. B.; Papayannakos, N. G., Group additive kinetic modeling for carbon-centered radical addition to oxygenates and -scission of oxygenates. *AIChE Journal* **2016,** 62, (3), 802-814.
59. Saeys, M.; Reyniers, M.-F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M., Ab initio group contribution method for activation energies for radical additions. *AIChE Journal* **2004,** 50, (2), 426-444.
60. Van de Vijver, R.; Sabbe, M. K.; Reyniers, M.-F.; Van Geem, K. M.; Marin, G. B., Ab initio derived group additivity model for intramolecular hydrogen abstraction reactions. *Physical Chemistry Chemical Physics* **2018,** 20, (16), 10877-10894.
61. Davis, A. C.; Francisco, J. S., Ab Initio Study of Hydrogen Migration across n-Alkyl Radicals. *The Journal of Physical Chemistry A* **2011,** 115, (14), 2966-2977.

62. Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H., Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications* **2016,** 203, 212-225.
63. Merchant, S. S. Molecules to Engines: Combustion Chemistry of Alcohols and their Applications to Advanced Engines. Massachusetts Institute of Technology, 2015.
64. Fannin, G., *Distillation Process Analyser with ASTM 86 Compliance*; 'Technical report by' Bartec Benke GmbH: 2013.
65. Ferris, A. M.; Rothamer, D. A., Methodology for the experimental measurement of vapor–liquid equilibrium distillation curves using a modified ASTM D86 setup. *Fuel* **2016,** 182, 467-479.

BEYOND STRUCTURE-PROPERTIES RELATIONSHIPS: TEMPORAL ANALYSIS OF PRODUCTS (TAP) AND CHEMICAL CALCULUS

John T. Gleaves

Mithra Technologies, Inc.
Foley, MO 63130, USA
j.gleaves@att.net

Catalytic properties of complex solids derive from nanoscale assemblies (active sites) of atoms and molecules working in-concert with the underlying bulk structure. Catalytic properties emerge during synthesis, activation, and reaction and are the result of a complex sequence of physical and chemical processes that are kinetically controlled. During a catalytic process, molecules interact with active sites to form products. Under reaction conditions diffusion and exchange of atoms, along with the exchange of energy, create a dynamic environment in which the composition and structure of the surface can change.

Catalyst development methodologies typically involve data-driven high-throughput testing of vast libraries of composition and surface science studies of model surfaces. This talk presents the concept of chemical calculus in which a researcher incrementally builds a new nanoscale chemical architecture and at every change in composition revises the direction of the experiment based on analysis of the evolution of kinetic properties. This approach addresses complexity by focusing on the evolution of kinetic properties to reveal connections between the physical features of the material and the multi-step reaction mechanism. A distinguishing feature of chemical calculus is a process called IKS (Incremental Kinetic Synthesis) that involves infinitesimal changes in surface composition, to understand the role of different components and processes in forming the nanostructures that control a complex reaction mechanism. Modified surfaces are fabricated on micron-sized particles using atoms, clusters and nanoparticles as component parts.

The modified particles are reactively characterized using TAP (Temporal Analysis of Products) pulse response experiments. Feedback from these experiments provides unique information, which describes the kinetic state of the sample. During TAP experiments the deposited components may react with probe molecules, may self-assemble into unique nanostructures or undergo phase transitions that induce changes in the kinetic dependencies that are uniquely observable with TAP. With the addition of more metal atoms the surface composition further evolves and this change will propagate through the reaction mechanism and be reflected in the kinetics. Each iterative cycle provides new insight into the nature of