Contents lists available at ScienceDirect

Engineering

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

Research
Smart Process Manufacturing: Deep Integration of AI and Process Manufacturing—Article

蒸汽裂解建模中的人工智能——详细流出物预测深度学习算法

Pieter P. Plehiers [a], Steffen H. Symoens [a], Ismaël Amghizar [a], Guy B. Marin [a], Christian V. Stevens [b], Kevin M. Van Geem [a,*]

[a] *Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Gent 9052, Belgium*
[b] *SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Gent 9000, Belgium*

ARTICLE INFO

Article history:
Received 14 September 2018
Revised 7 January 2019
Accepted 12 February 2019
Available online 9 October 2019

关键词
人工智能
深度学习
蒸汽裂解
人工神经网络

摘要

化工过程可以从快速准确的流出物成分预测中获益良多，以进行工厂设计、控制和优化。工业4.0革命宣称，通过将机器学习引入这些领域，就可实现可观的经济收益和环境收益。高频优化和工艺控制的瓶颈往往是按要求对原料和产品等进行详细分析所需的时间。为解决这些问题，已为最大的化工品生产工艺——蒸汽裂解建立由4个深度学习人工神经网络（deep learning artificial neural network, DL ANN）组成的框架。所提出的方法可根据有限数量的石脑油商业指标和可快速获得的工艺特性，来确定石脑油原料的详细特性和蒸汽裂解炉流出物的详细组成。根据沸点曲线上的三个点和PIONA（paraffin, iso-paraffin, olefin, naphthene, aronatic, 烷烃、异链烷烃、烯烃、环烷烃和芳香烃）特性预测石脑油的详细特性。若沸点不可用，则同时对沸点进行估计。即使在沸点是被估计的情况下，所建立的深度学习人工神经网络仍优于已有的香农信息熵最大化和传统人工神经网络等方法。对于原料重构，在测试集得到的平均绝对误差（mean absolute error, MAE）为0.3%（质量分数，下同），流出物预测的平均绝对误差为0.1%。结合所有网络使用前一网络的输出作为下一网络的输入——流出物平均绝对误差增大至0.19%。除这些网络具有高精度外，主要好处是获得预测值所需的计算成本可忽略不计。在标准的英特尔i7处理器上，预测值以毫秒为单位。COILSIM1D等商业软件在精度方面表现稍好一些，但每个反应所需的中央处理器时间以秒为单位。精度损失极小，而速度大大提高，使所提出的框架非常适用于工艺参数难以获得的连续监控过程，且非常适用于预想的高频实时优化（real-time optimization, RTO）策略或工艺控制。然而，由于这种方法缺乏基本依据，这意味着这种方法几乎丧失了可解释性，而这并不总是被工程界广泛接受。此外，对于那些训练集之外的，所建立网络的性能有明显下降。

1. 引言

无论是现在还是在可预见的将来[1]，大多数低碳烯烃是通过蒸汽裂解来生产的，所以利用此领域的新技术发展和创新很重要。人工智能（artificial intelligence, AI）就是其中一种在过去几年已经席卷全球的新技术。策略博弈[2,3]、自然语言处理[4,5]以及自动驾驶汽车[6,7]等多个领域已广泛采用AI。最近，AI技术已经进入化学（工程）研究领域[8]。AI也正缓慢而平稳地进入工业制造和生产过程[9]。固然，与汽车行业等相比，散装化学品行业在这一转变中相对保守。即将到来的被称为“工业4.0”的技术革命，有望重新定义生产极限[10–14]。AI在化学中的应用实例包括药物研发[15,16]与合成[17,18]以及计算化学[19]等。如上述实例所示，

* Corresponding author.
E-mail address: kevin.vangeem@ugent.be (K.M. Van Geem).

英文原文: Engineering 2019, 5(6): 1027–1040
引用本文: Pieter P. Plehiers, Steffen H. Symoens, Ismaël Amghizar, Guy B. Marin, Christian V. Stevens, Kevin M. Van Geem. Artificial Intelligence in Steam Cracking Modeling: A Deep Learning Algorithm for Detailed Effluent Prediction. *Engineering*, https://doi.org/10.1016/j.eng.2019.02.013

AI技术擅于处理高度复杂的、非线性的问题。因此，这些方法可以应用于蒸汽裂解工艺反应器段的建模，尽管这本身就是复杂且非线性的，但产生的模型有望在执行速度和精度方面优于传统的详细动力学模型。随着蒸汽裂解和其他行业[20−22]实时优化（real-time optimization, RTO）系统的复杂性不断增加，且性能不断提高，详细输入的必要性程度也日益提高。虽然这在技术上是可行的，然而因为其数据处理是劳动密集型且耗时的，目前工业领域仍未有成熟的方法使用综合、在线、二维气相色谱（two-dimensional gas chromatography, 2D-GC或GC × GC）进行详细蒸汽表征[23]。因此，RTO系统所需要的详细组成通常通过取样分析和离线分析获得。这些耗时的分析导致RTO系统每隔几小时只执行一个优化步骤[24]。以上并不意味着在线表征技术未应用于工业中；准确地说，与全面2D-GC相比，采用的在线表征技术所传达的详细信息往往要少得多。除了它们的实时优化值，反应器输入与输出组成的详细知识也对安全高效运行至关重要。此外，建立准确的反应器模型严重依赖于原料和流出物表征的细节层次。以上意味着原料重构和反应器建模算法是必不可少的。虽然这两个主题都不乏研究，但很少有方法涉及AI。Hudebine和Verstraete[25]、Verstraete等[26]以及后来的Van Geem等[27]都使用信息熵最大化方法在各种石油馏分的原料重构方面取得了巨大成功。因为动力学模型能够超出预定义的训练集范围进行推断，动力学模型在反应器建模中是占据主导地位的[28−35]。人工神经网络（artificial neural network, ANN）是一种常用的AI工具[36]。这种生物模拟形式是对人类大脑神经网络的一种简化数学表达，如图1所示[37]。

在原料重构方面使用AI的实例见Pyl等[38]的文章，他们建立的人工神经网络根据石脑油PIONA（paraffin, iso-paraffin, olefin, naphthene, aronatic，烷烃、异链烷烃、烯烃、环烷烃、芳香烃）组成和沸点曲线（boiling point, BP）确定裂解工艺中石脑油的详细分子组成。Niaei等[39]及后来Sedighi等[40]使用人工神经网络对反应器流出物组成进行了建模，但只有给定原料时才这样做。Ghadrdan等[41]在人工神经网络模型中引入一组9个原料类型的参数，以定性方式解决了这一原料缺口。传统人工神经网络和更经典的机器学习技术毫无疑问都是强大的工具，但都依赖开发人员识别对问题进行描述的正确特征。本文将深度学习（deep learning, DL）方法应用于石脑油原料的原料重构和反应器流出物预测问题。深度学习依靠网络本身，进一步利用人工神经网络的能力来识别输入、提取输入并将输入与包含更多问题解决（即预测输出）相关信息的抽象特征结合，如图2所示[42,43]。其想法是，这个新增抽象层次会提高网络泛化到不可见数据，并因此在网络训练集以外数据方面优于传统人工神经网络的能力。

接下来描述4个相互作用的深度学习人工神经网络，以达到对蒸汽裂解炉反应器流出物组成的预测精度为最终目标，使用有限数量的原料商业指标作为输入。图3说明了由相互作用的深度学习人工神经网络组成的框架。网络1使用最基本的输入——PIONA，以密度和蒸汽压作为输入来预测初沸点、中沸点和终沸点。网络2使用这些预测的沸点，结合之前指定的PIONA，对原料进行详细重构，然后将该重构用于网络3的输入。网

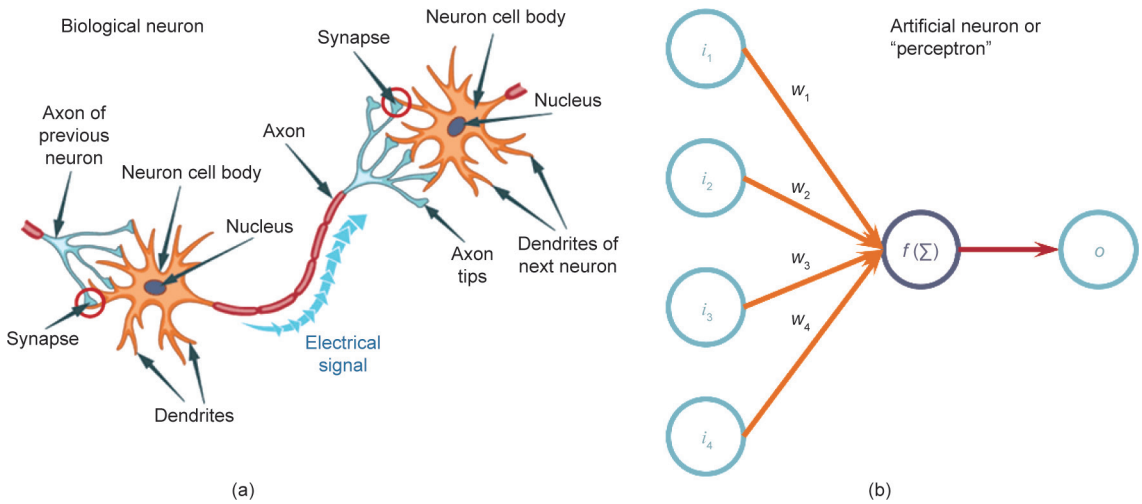


图1. 生物神经元（a）与人工神经元或感知器之间的类比（b），Mahanta [37]。i：输入；W：权值；o：输出；$f(\sum)$：激活函数。

络3预测流出物的详细组成。网络4用于网络1和网络2的延伸与检验。网络4使用石脑油的详细PIONA特性，预测石脑油的密度、蒸汽压以及前面提及的三个沸点。

在第3节介绍每个深度学习人工神经网络的架构前，我们在第2节简要论述了人工神经网络的理论，并在Supplementary data中给出了与数据相关的一些评论。第4节论述了网络经训练后的结果，并将这些结果与支持向量回归（support vector regression, SVR）和随机森林回归（random forest, RF）等其他重构方法和预测方法的结构进行对比。在最后一节，我们对蒸汽裂解流出物预测方法进行了简单总结和展望。

2. 方法与数据

2.1. 深度学习人工神经网络

深度学习人工神经网络的数学方面是相似的，因此，本节不再区分传统人工神经网络和深度学习人工神经网络[44]。式（1）给出了单个感知器的输入向量$\boldsymbol{i}$和输出o之间的关系。所有输入均按相应的权值w_j进行加权，然后求和。并且在加权和上加上一个常数偏置项b。激活函数f将非线性特征引入网络。常用的激活函数为sigmoid函数、双曲正切函数、线性整流函数（rectified linear unit, ReLU）函数和softmax函数。关于这些激活函数的更多信息，可见补充数据第S1.1节。单个感知器的方程式很容易扩展到式（2）来描述完整的网络层，其中，$\boldsymbol{W}$指该层的全矩阵。每个感知器可拥有自己的偏置参数。通过反复应用式（2），最终对整个网络进行数学描述，得出式（3），适用于含一个输入层、一个带偏置$\boldsymbol{b}_1$的隐藏层和一个带偏置$\boldsymbol{b}_2$的输出层$\boldsymbol{y}$的人工神经网络。

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

式中，$\boldsymbol{w}$为单个感知器的权向量；i为感知器的输入；j为

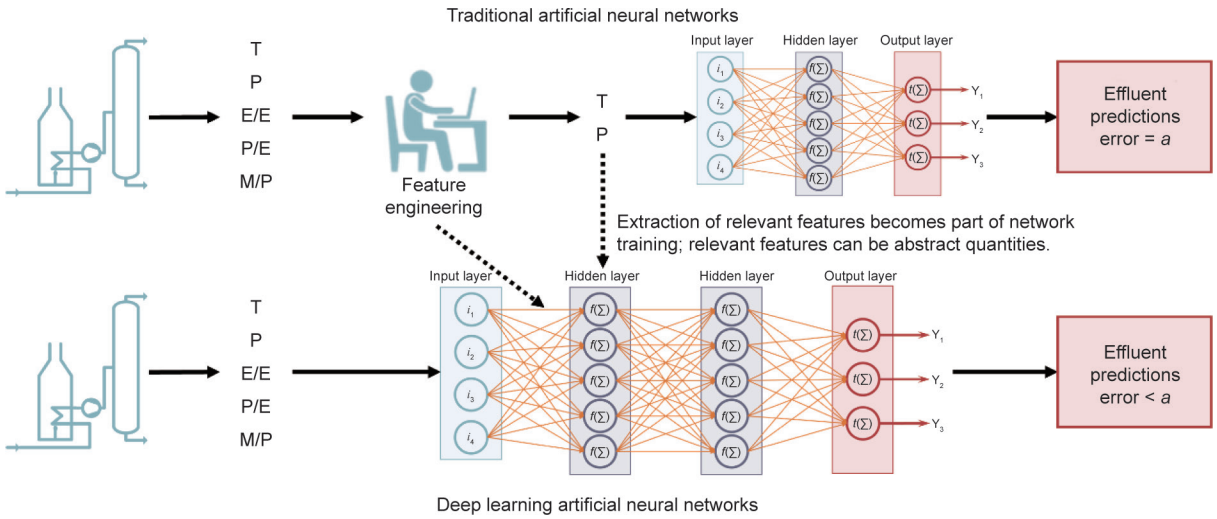


图2. 浅层人工神经网络与深度学习人工神经网络对比（Seif [43]）。T：温度；P：压力；E/E：产物乙烯与/乙烷比；P/E：产物丙烯/乙烯比；M/P：产物甲烷/丙烯比；Y_1：输出1；Y_2：输出2；Y_3：输出3；a：特定值；t (∑)：激活函数。

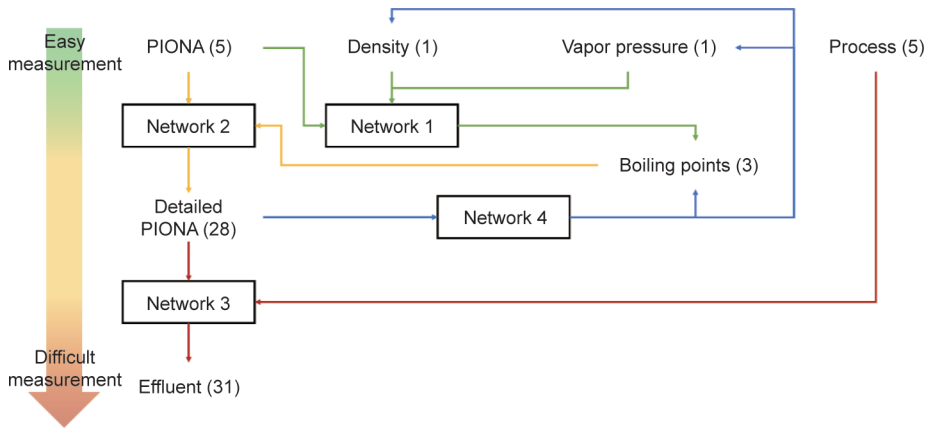


图3. 深度学习人工神经网络框架内不同变量和4个网络相互作用原理概观。括号中的数字表示每个变量描述符的数量。

层内的节点索引。

$$\boldsymbol{o} = f(\boldsymbol{W} \cdot \boldsymbol{i} + b) \quad (2)$$

式中，$\boldsymbol{o}$为层输出向量。

$$\boldsymbol{y} = \boldsymbol{f}_2[\boldsymbol{W}_2 \cdot \boldsymbol{f}_1(\boldsymbol{W}_1 \cdot \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2] \quad (3)$$

式中，$\boldsymbol{y}$为模型输出向量；$\boldsymbol{x}$为模型输入向量；$\boldsymbol{f}_1$和$\boldsymbol{f}_2$为第1层和第2层的激活函数；$\boldsymbol{b}_1$和$\boldsymbol{b}_2$为第1层和第2层的偏置向量。

本文中人工神经网络是通过反向传播算法[44,45]进行训练的，可从输出开始，将误差层层传递，实现对网络层权值的更新。用梯度下降优化算法使某个目标函数最小化。目标函数中频繁使用的误差度量方法为均方（根）偏差（root mean squared deviation, RMSD）、平均绝对误差（mean absolute error, MAE）和平均绝对百分误差（mean absolute percentage error, MAPE）。为了优化权值，通常需要对完整训练集进行几次迭代（iteration）。一次这样的迭代称为一个阶段（epoch）。在一个阶段内，训练集再分为几个批（batch）。网络权值每批更新一次。小批量（batch size）即每个优化步骤的样本的有限数量，从训练速度的角度出发，小批量会减少训练所需的阶段数目，但是会增大每个阶段的计算时间。而且，更小的批量会导致梯度估计更差，降低优化的稳定性。

在人工神经网络中，可在网络的过度拟合和过度训练之间做区分[46]。当网络太复杂时，即网络层数过多或每层节点数过多时，就会发生过度拟合。根据通用近似定理，对于任何函数，都可找到相对应的人工神经网络，使其在数据层面上以任意期望的精度逼近该函数[47]。另一方面，过度训练则与训练阶段数有关。若训练数据太频繁地出现在网络上，则网络将开始“记忆”数据；换言之，网络将试图预测准确的输出值，而不是数据泛化趋势中预期的值。这可用一个简单的例子说明。假设两个变量为线性相关。在数据集中，一个数据点由于测量误差等原因不遵循这一线性趋势。经过几个训练阶段后，网络已能认识线性趋势。然而，因为存在不遵循线性趋势的数据点，平方和仍然较高。在训练期间，平方和是以最小化的趋势演变的。因此，在后续每个阶段，网络将开始描述越来越不线性的趋势，因为在见过不遵循线性趋势的数据点多次后，网络“认为”此数据点也遵循线性趋势。过度训练可通过监测训练数据集和验证数据集的目标函数或网络精度来确定。尽管对于训练集而言，目标函数通常随着阶段数的增加而呈递减趋势，但验证数据的目标函数将在某个点开始再次递增。自这点起，网络就在被过度训练。以上问题是可以补救的，例如，在训练期间使用丢弃法（dropout）进行补救[48,49]。在丢弃法中，在每批数据训练期间，随机选择部分网络节点并将其暂时从网络中除去。这样，每个神经元须单独学习特征，不能依赖邻近的神经元捕获信息。本文所有网络的丢弃率为0.5。因为每步（step）中只更新一半权值，用丢弃法减少过度拟合的代价就是降低了网络学习速度。因为构建的网络在测试数据上表现良好，所以本文未评价目标函数的L1和L2标准化[50]等其他标准化方法。

本文使用Python深度学习库Keras [51]训练人工神经网络，该数据库采用Tensorflow后端[52]并配置图形处理器（graphics processing unit, GPU）加速。

2.2. 数据分析

2.2.1. 石脑油

Pyl等[38]的文章提供了272种详细的工业石脑油成分。可用的石脑油性质包括密度、蒸汽压、通过ASTM D86标准方法确定的三个沸点，即初沸点（IBP）、中沸点（BP50）和终沸点（FBP），以及每个碳数的详细PIONA分数。图S13（见Appendix A）提供了可用数据的相关矩阵。从图中可观察到，蒸汽压和初沸点是强相关的，正如密度和中沸点一样。终沸点与密度和蒸汽压的相关性不那么强，但是与中沸点存在明显的相关性。这种相关性将影响根据石脑油蒸汽压和密度预测沸点的网络的架构，详情见第3.1节。

本文与Pyl等[38]的文章一样，根据数据集的10个输入变量（初沸点、中沸点、终沸点、密度、蒸汽压和PIONA）进行了主成分分析（principal component analysis, PCA）[53]（详情见Appendix A中的第S1.2节）。图4展示了主成分分析结果。从图4（a）可以得出的结论是：（训练）数据集由三个成分描述。根据图4（b）所示的这些主成分（principal component, PC）中前两个成分的输入值得分，我们就可以确认相关分析的结果。例如，在密度和中沸点之间观察到的高度相关性转换成主成分空间中的平行向量。尽管初沸点和蒸汽压方向相反，但它们表现出类似特性。

在主成分分析基础上对测试集进行第二次分析。由于人工神经网络在训练期间只依赖于训练数据集和验证数据集，可以预见的是，只有与训练数据和验证数据相似的测试数据才会产生准确结果。确定某数据点与训练

集是否相似的一个衡量标准是马氏距离（Mahalanobis distance, MD）[54,55]。在主成分空间，可用式（4）计算马氏距离。

$$MD^2 = \boldsymbol{z}^{T} \cdot (\boldsymbol{\Lambda}')^{-1} \cdot \boldsymbol{z} \quad (4)$$

式中，$\boldsymbol{z}$为主成分空间内的输入，且包含每个主成分上原始输入的得分。$\boldsymbol{\Lambda}$为所有特征值的矩阵，在本文中为10 × 10对角矩阵。$\boldsymbol{\Lambda}'$为约化的3 × 3特征值矩阵，只包含与所选三个主成分对应的特征值。马氏距离较大的石脑油成分可被视为异常值，因此预计预测情况较差。图4（c）和（d）所示为主成分空间内的测试集分布。虚线对应的是2.5的马氏距离，表示椭圆体内的石脑油在训练集范围内的概率为90%。马氏距离取值为2.5，用作考虑相应的石脑油是否为异常值的临界距离。有一个石脑油成分（图4中用红色表示）的马氏距离为5.08。综上所述，分析结果表明，预测情况总体良好，但是也不排除对上述石脑油预测较差的情况。

2.2.2. 流出物组成

由于获取详细的工业蒸汽裂解炉流出物组成受到严格限制，因此，Van Geem等[30,56]和Vervust等[57]使用最先进的反应器模拟软件工具获取需要的流出物特性。COILSIM1D已经过大量专有数据验证，在工业中用于详细的蒸汽裂解炉模拟；因此，它是可靠、准确的工具，所获得的结果足以代替不可用的实验或工业数据。这种使用模拟数据代替不可用和（或）有限的实验数据的方法，在其他领域，尤其是在预测分子热力学性质和反应动力学时，已成为惯例[58–63]。使用模拟数据作为训练数据、获取实验数据的难度以及对准确输入输出数据的需要，既强调了详细的基本模型对模拟和了解这些过程的持续重要性，又强调了高精度实验方法的关键必要性。

COILSIM1D可预测输出量中数百种不同的化学物质。然而，这些成分的大部分对蒸汽裂解炉的整体运行不大重要。因此，有28个（伪）成分得到了确认。这些包括乙烯、丙烯、苯、氢和丁二烯等若干分子组分和C_7异链烷烃、C_{10+}芳香烃等集总组分。完整的成分清单见第S2节。我们一共运行了两组模拟实验。

第一组共包含13 600次模拟，用于训练和测试网络对详细流出物组成进行预测。每次模拟使用一个不同的石脑油成分。这些不同的石脑油成分来自2.2.1节所述数据集，但在各浓度中引入了0~10%的随机变化。每一个石脑油成分都与一系列不同的工艺条件相结合。这些工艺条件为炉管出口压力（coil outlet pressure, COP）和炉管出口温度（coil outlet temperature, COT）。图S14表

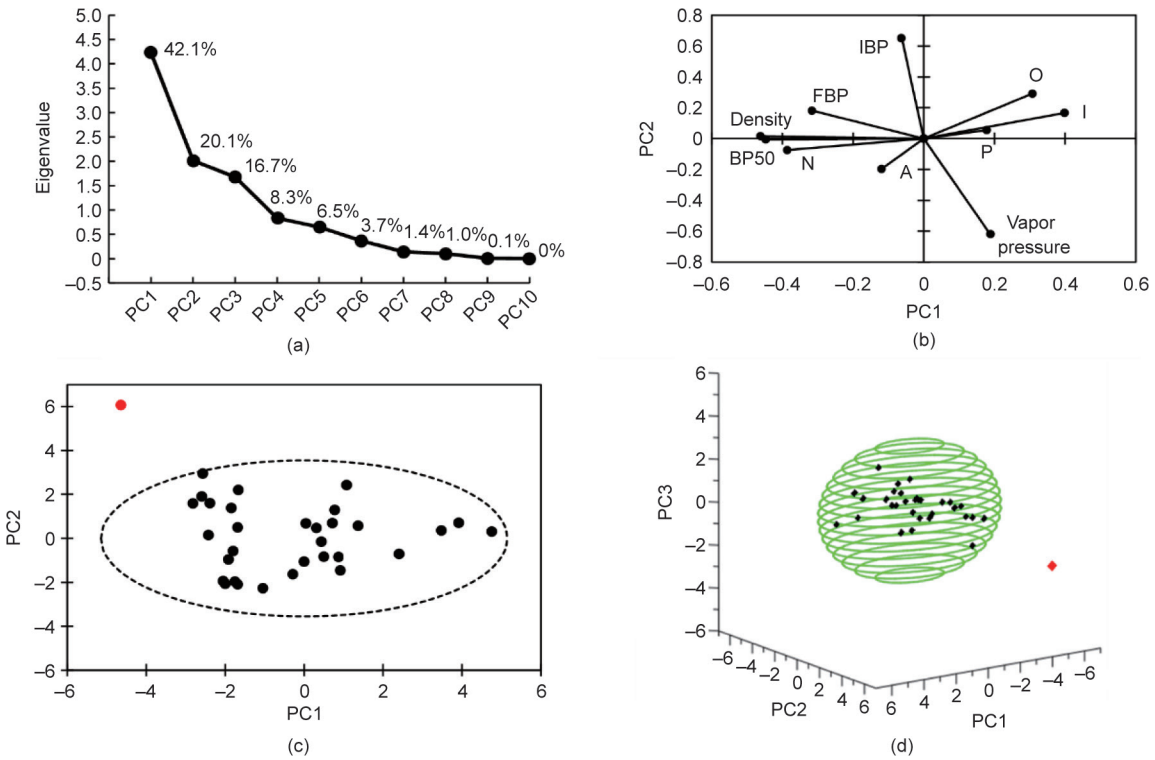


图4.（a）主成分的特征值和解释方差（数据点上方的百分数）；（b）沿着第一个主成分和第二个主成分的输入分解（得分图）；（c）、（d）石脑油测试集的主成分表示，异常值用红色表示。P：烷烃；I：异链烷烃；O：烯烃；N：环烷烃；A：芳香烃。

明，石脑油成分和工艺条件都以均匀的方式覆盖了大范围的变量空间。本文使用单个反应器和炉配置完成所有的模拟实验。后面会提到，确切的反应器配置不太重要。对新数据集进行主成分分析，识别可能有问题的情况。图5（a）表明，数据集由6个主成分描述。将测试数据集投射到主成分空间的前三个维度时，如图5（b）所示，观察到有少量的输入位于椭圆之外，椭圆包含90%的训练数据，且对应的马氏距离为3.3。同样，这表明测试集上的整体性能都是良好的，只有在极个别的情况下预测性能不佳。

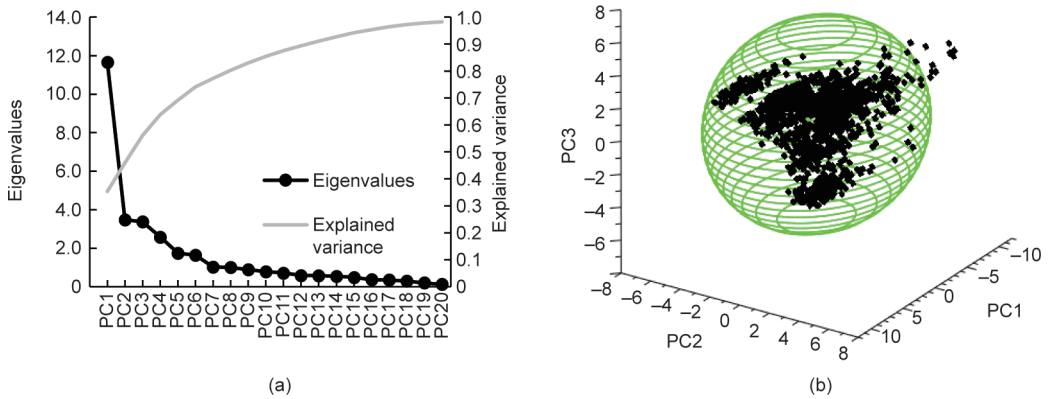
第二组包括1587次额外模拟，用于测试完整工作流程和各网络的综合性能。使用与前面的模拟相同的反应器和炉配置。这组共考虑32个石脑油成分，对应于网络1、网络2和网络4的测试集，因此在测试期间未使用任何训练数据。每一种石脑油都由一组固定时间间隔的工艺条件来扩展。在750~950 ℃的范围内，考虑10个炉管出口温度。同样，在1.7~2.3 bar（1 bar=10^5 Pa）范围内考虑5个炉管出口压力。尽管这导致对变量空间的覆盖是类似网格的，但也足以达到测试目的。

3. 人工神经网络的设置

3.1. 从密度和蒸汽压到沸点

本文旨在建立一组算法，让用户只使用现成的描述符就能够获得对蒸汽裂解反应器流出物的详细预测。由于使用详细原料特性时详细预测才更可靠，所以该算法的第一步是根据原料的商业描述符重构原料。根据Van Geem及其同事以往的文章[27,30−32,38]，很明显，至少需要石脑油沸点曲线上的某些点才能成功重构石脑油组成。然而，沸点难以在线测量，因为单次符合ASTM 86要求的测量可能要花30~45 min [64]。因此，沸点不是现成的。所以，对沸点曲线上的三个重要沸点进行估计是预测流出物组成的第一步。沸点的预测以石脑油的密度、蒸汽压和基本PIONA特性为基础。根据2.2.1节所述原料参数之间的关联，本文构建了一个网络，其架构如图6所示。由于初沸点和中沸点均与终沸点有很强的相关性，所以本文将包含初沸点和中沸点估计值的向量与第一隐藏层连接。这样，网络在预测终沸点期间就能直接使用初沸点和中沸点的预测值。在输入层上选择第一隐藏层，是因为在深度学习法中，网络会学习第一隐藏层中与预测输出量最相关的输入表示。此后，该网络将被称为网络1。为提高网络的稳定性和性能，将所有输入和输出均标准化到数据集的范畴内。表1列出了每个变量标准化后的最大值和最小值。根据80∶8∶12的比例，将由272个石脑油成分组成的数据集分成训练集、验证集和测试集。验证集用于调整网络的超参数，在本研究中，超参数包括各隐藏层中的节点数、批量、激活函数以及训练阶段数。一般而言，“超参数”一词表示除节点权重和偏置量之外的所有网络参数。本文以启发方法搜索最优组合。关于此搜索的更多详细信息，请见S3.1节。测试集用于评价最终优化的网络。

图6展示了得到的超参数及其架构。比较不同超参数下的网络性能的其他图，可参见S3.2节。鉴于在考虑的超参数网格（S3.1节）下，最后选择的网络具有较低的均方误差，本文首选平均绝对误差作为训练目标函数，而不是均方误差。关于此特定网络的详细解释见S3.2节。由于独立成分的标准化，所有输出量均具有类似的数量级。因此，使用平均绝对百分误差不利于网络精度。就平均绝对误差而言，批量为8时，经过1181个训练阶段后，就能达到最佳性能。在训练数据和验证数据上对使用优化超参数的最终网络进行训练，之后根据不可见测试数据对网络进行验证。



(a) (b)

图5.（a）流出物数据集主成分分析中前20个主成分的特征值和解释方差；（b）降至三维的主成分空间内的流出物测试数据。

3.2. 原料重构

框架中第二个网络使用石脑油的PIONA组成和沸点来重构原料的详细组成。在训练该网络时，使用实验沸点作为输入。根据Pyl等[38]的文章，本文估计了28个与图S15中的详细PIONA矩阵对应的不同的伪成分（pseudo-component）。另外，分别在A_8类别和N_6类别中对二甲苯和乙苯以及对环己烷和甲基环戊烷进行了区分。按前一节中的程序和表1所列范围将输入标准化。对于输出，应用了不同的标准化程序。不同类别中各成分的绝对浓度范围很广。C_5和C_6的质量分数高达35%，而烯烃的质量分数可降至0.01%。如果试图使用单个softmax函数来直接一次性预测所有分数，那么将导致网络难以训练，尤其考虑到可用的训练数据量有限，更是如此。使用单个softmax函数层的好处在于：输出总和为1，与期望质量分数的物理性质对应。但是，由于质量分数范围广，输出中5个PIONA类别是按照式（5）中的烷烃示例分别标准化的，式（5）如下：

$$P_i^{\mathrm{norm}} = \frac{P_i}{\sum_j P_j} \tag{5}$$

式中，P_i^{norm}为标准化的PIONA类别；P_j为一个PIONA类别。

首先将输出层分为5个独立的输出，除烯烃质量分数之外的每个单独的成分类别可使用一个softmax激活函数。由于总烯烃浓度可能为零，根据softmax激活函数的性质，网络被迫错误预测了总和为一的烯烃分布。这对总体精度有不利影响。因此，对于烯烃输出层，使用sigmoid激活函数。产生的多输出架构和优化的超参数如图7所示。接下来，该网络被称为网络2。再次按80∶8∶12的比例对数据进行训练/验证/测试划分。补充数据S3.3节介绍了关于优化的更多细节。简言之，选择平均绝对误差作为网络目标函数。由于按成分类别对输出进行了标准化，所以输出未跨越几个数量级，因此无需相对费用函数。对于网络2，使用16个批量和45 285个训练阶段，就能达到最佳性能。

表1 网络1的输入变量和输出变量范围

Variable	Minimum value	Maximum value
IBP (K)	303	328
BP50 (K)	323	398
FBP (K)	348	463
Denisity	0.65	0.75
Vapor pressure (kPa)	27.6	84.9
Paraffins (wt%)	27.5	50.0
Isoparaffins (wt%)	25.0	52.5
Olefins (wt%)	0	1
Naphthenes (wt%)	5	35
Aromatics (wt%)	0	17

3.3. 详细流出物预测

第三个网络以详细PIONA组成（28个伪成分）和5个工艺特征作为输入，预测蒸汽裂解炉反应器流出物的详细分子组成。如2.2节所述，此网络使用经调整的

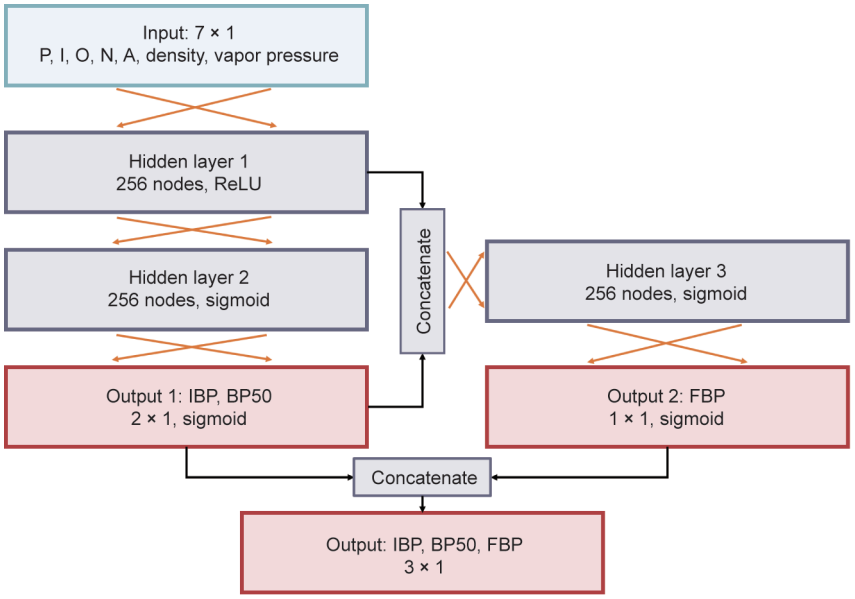


图6. 网络1的架构——根据石脑油PIONA组成、蒸汽压和密度预测初沸点、中沸点和终沸点。

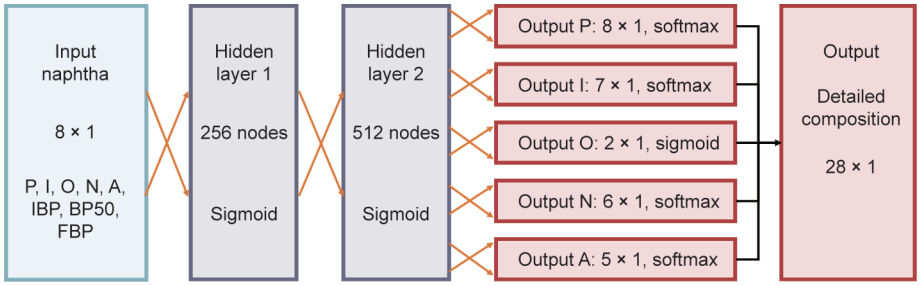


图7. 网络2的架构——以PIONA特性和沸点开始重构更详细的原料组成。

数据集，此数据集包含的数据点比前两个网络所用的数据集多50倍。详细PIONA组成中考虑的成分与图S15中所示相同。根据Van Geem等[31]以往的研究，确定了5个工艺描述符。前两个描述符即炉管出口温度和炉管出口压力，已用于产生数据集。剩下三个描述符为产物乙烯/乙烷比（ethylene to ethane, E/E）、产物丙烯/乙烯比（propylene to ethylene, P/E）和产物甲烷/丙烯比（methane to propylene, M/P）。Van Geem等[31]的研究已证明，对于指定的石脑油，流出物组成完全是由其中两个描述符确定的。然而，如图S16所示，当这5个描述符都作为输入时，获得的模型会更准确。准确度提高有三方面的原因。首先，因为甲烷、乙烷和丙烯的质量分数是能够根据预测的乙烯质量分数计算出的，所以使用上述产物比作为输入，模型所需预测的输出就会减少三个。其次，因为不确定性是散布在多个输入上的，所以当引入本质上描述温度和压力的相同工艺参数的多描述符时，模型就具有对输出误差的鲁棒性。第三个原因也是最重要的原因，可以追溯到深度学习网络的能力上，如图2所示。通过训练多个输入上的多层网络，深度学习网络就能够从诸多输入中自由提取其认为与解决流出物组成预测问题最相关的信息。如果仅使用炉管出口温度和炉管出口压力来训练模型，那么深度学习的潜力就得不到充分利用。如果手动选择或设计网络输入并从网络输入中去除某些工艺描述符，那么网络上绝不会显示数据中潜在有用的信息。总之，这5个确定的描述符都要作为网络输入。

在表2给出的范围内对炉管出口温度、炉管出口压力、乙烯/乙烷比、丙烯/乙烯比和甲烷/丙烯比的值进行标准化。由于输入值之间的尺寸不匹配，所以第一层分为工艺和原料特征层，产生的深度学习人工神经网络比常规紧密相连的人工神经网络更先进。这种划分允许从两个输入中提取的特征向量是独立、等长、相关的。由于这不是网络预测的完整流出物频谱，所以输出的总和不宜等于1。因此，softmax激活函数不可应用于输出层，相反，利用sigmoid激活函数，考虑成分分数以0和1为界。最终的架构和超参数如图8所示。在这种情况下，选择平均绝对百分误差作为目标函数。关于这种选择的理由，详见S3.4节。本文中进一步应用此网络作为网络3。对于该数据集，应用训练/验证/测试比为81∶9∶10。批量为8，训练阶段为2744时，此网络就能达到最佳性能。关于超参数优化的更多信息，参见S3.4节。

表2 网络3工艺相关输入变量的范围

Variable	Minimum value	Maximum value
COT (K)	948	1318
COP (bara)	1.36	2.74
E/E	2	37
P/E	0	1.4
M/P	0	35

3.4. 性质估计

此框架中的最后一个网络用作对前两个网络的检验。根据详细石脑油组成，对密度、蒸汽压、初沸点、中沸点和终沸点进行估算。数据集与网络1和网络2用于反向运算的数据集相同。如果重构准确，那么重构石脑油的预测性质应该与真实石脑油的预测性质相差不大。有人可能认为，同时优化4个网络才可获得最佳结果。然而，考虑到数据集的大小有限，用多个反馈回路训练如此复杂的网络，在最坏情况下是不可行的，在最好的情况下，既不准确又不具有泛化性。第四个网络即网络4，具有简单的两层架构，有28个输入和5个输出，如图9所示，并有优化的参数。与网络1和网络2的原因类似，网络4也选择平均绝对误差作为损失函数。28个输入与重构算法中考虑的成分相同，如图S15所列。28个输入的总和标准化为1，输出根据表1所列的相同范围进行标准化。批量为8，训练阶段为5385时，网络4就能达到最佳性能。关于优化的更多信息，见S3.5节。

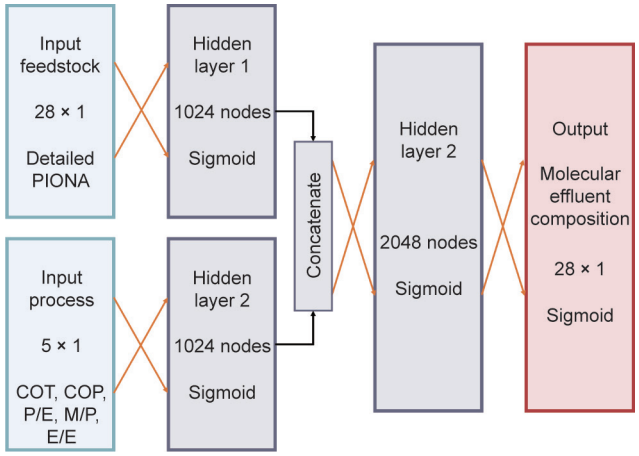


图8. 网络3的架构——根据详细原料组成和5个工艺描述符预测流出物的分子组成。

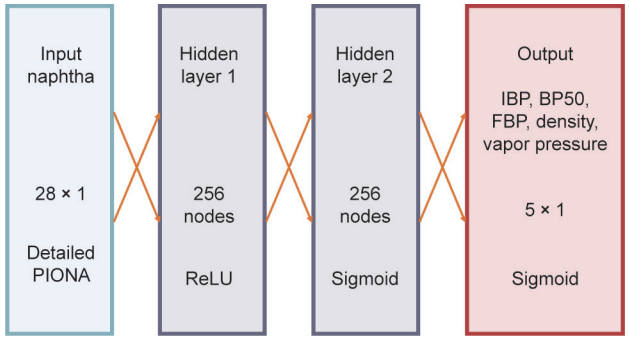


图9. 网络4的架构——根据详细PIONA特性预测石脑油性质。

4. 结果与讨论

4.1. 原料

网络预测初沸点、中沸点和终沸点的性能，如图10所示。总的说来，网络表现非常好，每个不同石脑油成分只有两个明显较差的预测。这些较差的预测值在图10中用红色和绿色表示。对于用红色表示的预测值，为其计算的马氏距离为1.82，低于2.5的临界值（第2.2.1节），所以预测应该是准确的。误差较大的原因将进一步讨论。绿色预测值对应的石脑油成分的马氏距离为5.08，相应地，对于自由度为（3, 237）的F-统计量，该成分属于训练集的假设成立的概率为2×10^{-5}。初沸点的预测值较差，超出了表1所示的输出标准化范围，表明网络必须预测大于1的值，但这是不可能通过构建网络实现的。然而，对于另外两个沸点，尽管石脑油成分与训练数据集有很大不同，但模型的预测非常准确。另外三个石脑油成分的马氏距离均大于2.5的阈值。这些石脑油成分的预测值与实验值偏离高达10 K。这表明了深度学习或任何其他类型回归的一个缺陷：如果输入值与

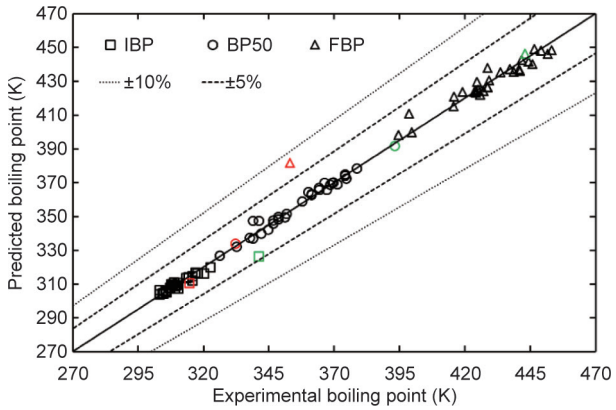


图10. 网络1的奇偶图（parity plot）；根据PIONA、密度和蒸汽压预测初沸点、中沸点和终沸点。用绿色表示的点是马氏距离为5.08的石脑油预测值；用红色表示的点是马氏距离为1.82的石脑油预测值。

训练集中的输入值有很大不同，那么有可能导致预测值较差。

表3[27]显示，所有沸点的预测值平均偏离实验值约1%或3 K。这一发现进一步证实：不必考虑以平均绝对百分误差为指标训练网络。网络精度与实验方法的精度不太匹配，例如，Ferris和Rothamer [65]报道的最大平均绝对误差为（2.2 ± 1.4）K的实验方法精度。然而，深度学习人工神经网络确实比Van Geem等使用的香农信息熵最大化（maximization of Shannon entropy, MSE）方法[27]表现更好。这种观察结果并不出乎意料。使用的大部分测试集，在训练期间并未被网络见到位于马氏距离为2.5或概率水平为0.9所对应的椭圆体范围内。因此，即使在测试集上也可以预见网络具有良好的性能。甚至对于临界椭圆体外部的数据点，深度学习人工神经网络模型的性能仍然能保证具有与香农信息熵最大化方法的相似性能。这点可由它们类似的最大偏差证明。

用深度学习人工神经网络能达到非常高的通量：在2.7 GHz英特尔i7-6820HQ中央处理器上预测32个测试石脑油成分的沸点耗时137 ms，或者说每个石脑油成分的沸点预测仅耗时4 ms多。不幸的是，使用Van Geem等[27]的方法无法进行相同速度的测试，因为沸点的估算属于原料重构的一部分；不过，考虑到总时间为25 s，可以认为深度学习人工神经网络会更快。

图11所示为网络2在所选数量的输出成分上的性能。所有输出成分的奇偶图见图S17。一般而言，在整个浓度范围内，网络的性能都良好。网络的总体平均绝对误差达到0.31%。两个边远的预测值均用红色挑出来。在2.2.1节提到，I_7成分与任何其他变量之间均缺乏相关性。当排除与红色高亮点对应的石脑油成分时，就能发现I_7成分组与其他变量之间的相关性增加了1%以上。

表3 与Van Geem等的研究[27]相比，网络1在测试集上的性能统计指标

Variable	MAE		RMSD (K)		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

由于排除的数据占数据的比例约为0.7%，所以可得出的结论是，这些数据对缺乏相关性具有重大影响。为石脑油成分A和石脑油成分B计算的马氏距离分别为2.27和1.82。因此，没有迹象表明石脑油成分位于训练集范围之外。上述研究表明，测量误差有可能引起预测值较差。这种可能性可由以下事实得到进一步证明：P_4和P_7等其他成分偏离趋势的预测值几乎全是由这两种有问题的石脑油成分导致的。一个或多个成分的测量误差也有助于解释图10中用红色高亮显示的石脑油成分的终沸点预测值为何较差，因为该石脑油成分与石脑油成分B相同。这一结果强调高质量输入对准确训练网络和获得准确预测值都非常重要。

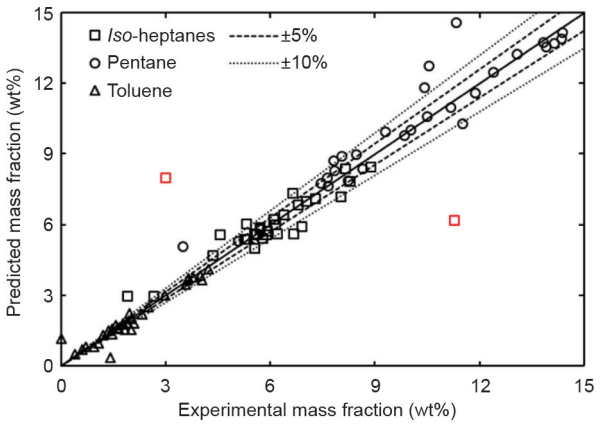


图11. 网络2在所选输出成分上的性能。

将网络2的性能与Van Geem等[27]和Pyl等[38]以往关于原料重构的研究对比，并与两个额外构建的模型相比；重构算法以下列方法为基础：香农信息熵最大化（Van Geem等[27]）、多元线性回归（multiple linear regression, MLR）（Pyl等[38]）、传统人工神经网络（Pyl等[38]）、支持向量回归（SVR）和随机森林（RF）回归法。将属于传统方法的多元线性回归法用作性能基线。表4所示为不同模型在各输出成分上的性能。与多元线性回归和香农信息熵最大化等更传统的方法相比，支持向量回归和（深度）人工神经网络等机器学习法有明显改进。图12所示为在平均绝对误差方面各模型的相对性能。深度学习法明显优于所有其他模型：网络2获得的平均绝对误差仅超过多元线性回归平均绝对误差的一半，且仍然比人工神经网络的平均绝对误差低20%。即使使用以密度和蒸汽压为基础的预测沸点，结合网络1和网络2，深度学习人工学习网络仍然明显优于所有其他测试模型。尽管香农信息熵最大化法的平均绝对误差更高，其优势在于：其依赖的是逐案优化，即此方法的适用性不局限于某一训练集的范围。就需要的中央处理器时间而言，香农信息熵最大化法需要大约25 s来模拟两个沸点和重构测试集的详细组成。使用网络1和网络2，合并后的进程在上述英特尔i7处理器上需要的时间仅为其十分之一，即234 ms。

网络4也与原料相关，因为网络4以已知的详细组成为基础来估计性质。网络4的性能用图13中的奇偶图说明。图13（a）中挑选出的预测值对应上述石脑油成分B的预测值。此外，蒸汽压预测较差可能是石脑油组成分析期间的测量误差导致的。表5所示为网络性能的统计数据。表5中也显示了网络1、2和4的组合性能。从最基本的商业指标开始时，网络的性能有明显降低；然而，仍然能获得合理准确的结果，总体性质趋势的预测仍然很好。

4.2. 流出物

由于使用了不同的训练和测试集，所以首先单独评估网络3的性能。为了保持图标的清晰度，下面所有的图都是通过从测试集中1360个数据点随机抽取其中10%得到的。对整个测试集计算统计指标。图14说明了在所选4个输出成分（即乙烯、1,3-丁二烯、氢和A_{10+}伪成分）上的网络性能。所有其他成分的奇偶图见图S18，此图表明网络在其他两个主要裂解产物（甲烷和丙烯）方面的性能与在乙烯方面的性能非常相似，如图14（a）所示。网络在整个质量分数范围内表现良好。对于乙烯、丁二烯和氢，质量分数范围被限制在一个数量级左右。但是，对于A_{10+}伪成分，数据集的质量分数分布在近4个数量级上。通过在几个数量级上准确预测A_{10+}伪成分的质量分数，网络证明了其预测能力。表6列出了这4个成分的具体统计数据以及所有成分的平均值。总之，该网络

表4 以PIONA和沸点为基础进行石脑油详细重构的不同算法的平均绝对误差（质量分数，%）

Component	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_{11}	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

MBP: modeled boiling point.

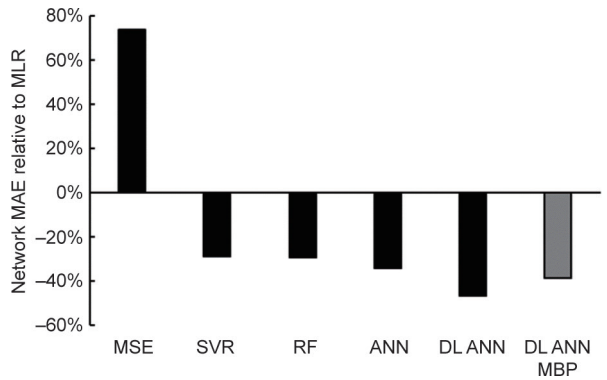


图12. 相对于多元线性回归模型平均绝对误差的网络平均绝对误差。深度学习人工神经网络模拟沸点使用模拟沸点作为输入，并结合网络1和网络2的性能。

的精度达到了0.1%，考虑到预测的最小计算成本，其精度是非常高的。还是在标准的英特尔i7笔记本电脑中央处理器上预测整个测试集的1360个反应，预测时间为1.716 s，或者每次预测的时间仅为1.2 ms。最先进的工具COILSIM1D需要数秒才能确定单个石脑油成分的详细流出物组成，这表明深度学习人工神经网络模型的速度大大提高了。几乎可以忽略的计算时间，允许在更大的、向流程提供反馈的频率比当前实时优化算法更高的实时优化算法中使用这样的网络。以这样的计算速度，甚至连前馈过程控制应用都是可能的。但是，速度提高带来的主要好处是：用有限的输入就能够持续监测难以获取的工艺参数，这有助于预测可能对下游运行参数重大（安全）影响的突变。

2.2.2节提到，确切的反应器配置不太重要。Van Geem等[31]已证明，对于指定的石脑油，其反应器流出物组成是由出口压力和温度的两个强度指数确定

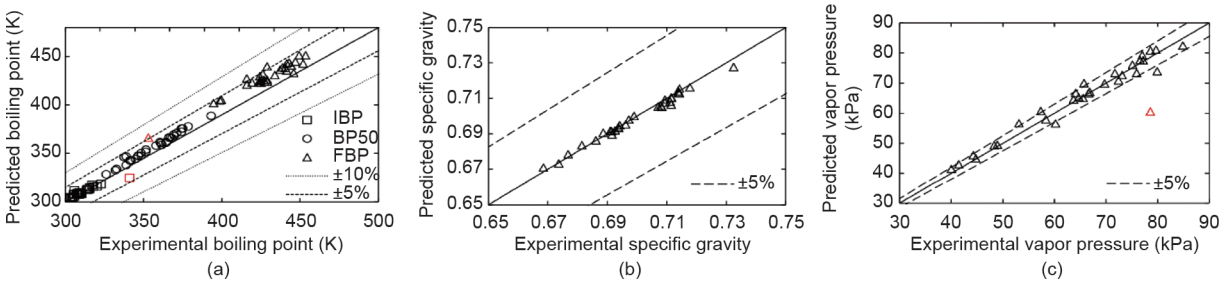


图13. 网络4不同输出的奇偶图。（a）初沸点、中沸点、终沸点;（b）密度作为比重;（c）蒸汽压。红色数据点对应石脑油成分B。

表5 以石脑油的蒸汽压和密度为基础，在测试集上和在测试集重构上网络4的性能统计

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

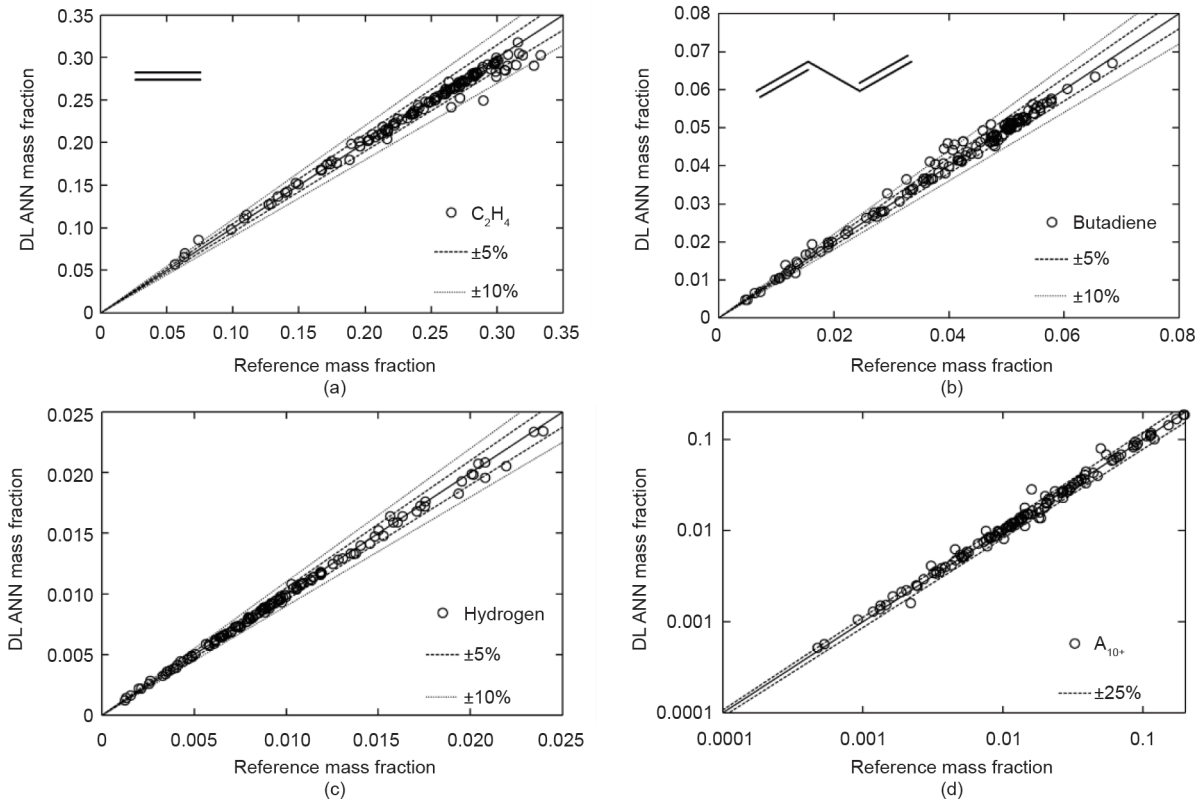


图14. 网络3对所选4个成分的预测结果奇偶图。（a）乙烯（C_2H_4）;（b）丁二烯;（c）氢;（d）伪成分。在1360个测试集数据点中，显示了136个点。

的，与反应器几何结构无关。网络3使用这些强度指数（即丙烯/乙烯比和乙烯/乙烷比）作为输入。因此，网络的性能与反应器几何结构相对独立，对于任何类型的反应器都能获得好的预测结果。这些发现的图示证明见图S19。

4.3. 流出物组合预测性能

最后，把利用简单、快速可得的指标进行的原料重构与详细流出物预测组合起来，评估组合性能。这相当于评估图3所示框架的性能。

运行组合框架的计算成本仍然很低。对1587个测试

表6 测试集上网络3在所选成分上的性能统计

Component	MAE (wt%)	MAPE (%)	RMSD (wt%)
Ethylene	0.42	1.9	0.763
Butadiene	0.10	3.1	0.150
Hydrogen	0.02	1.8	0.029
A_{10+} pseudo-component	0.18	7.3	0.762
Average	0.13	7.3	0.416

表7 测试集上网络1、2和3在所选成分上的组合性能统计

Component	MAE (wt%)	MAPE (%)	RMSD (wt%)
Ethylene	0.46	1.9	0.594
Butadiene	0.16	3.9	0.206
Hydrogen	0.02	3.2	0.030
A_{10+} pseudo-component	0.95	35.1	1.167
A_{7+} pseudo-component	0.43	8.9	0.594
Average	0.19	15.0	0.385

用例进行模拟的时间不到3.25 s，每个反应器的模拟时间为2 ms，只比利用详细石脑油特性模拟流出物所需时间多一点点。这表明，组合框架至少在计算方面适用于实时优化算法中或者甚至直接工艺控制中的融合。

通过对比图14和图S20以及对比表6和表7，观察到网络1、2、3的组合性能有下降。对于乙烯、丁二烯和氢等成分，网络精度仍然很高，接近使用真实石脑油组成达到的精度。网络在正确预测A_{7-9}和A_{10+}之间的分布方面确实有很大的困难。A_{7-9}的奇偶图见图S21，A_{10+}的奇偶图见图S20（d）。轻芳烃的浓度始终是被高估了，而重芳烃的浓度始终是被低估了。将这两个伪成分再次归为单个A_{7+}成分时，网络达到的精度类似于其他方法达到的精度，如表7倒数第二行所示。造成这种偏差的原因可能是非常轻微、系统性地低估了原料重构中具有较高浓度的芳烃。据观察，原料中芳烃含量发生微小变化，就可能对裂解过程中重芳香族化合物的形成产生显著影响。这表明了非常准确的实验数据的重要性，因为微小的测量误差就可能对结果产生重大影响。

奇偶图S20和图S21中的结果聚类是输入中网格状变化的结果。尽管工艺条件会影响输出物的准确特性，但对流出物组成产生主要影响的是石脑油组成。由于此数据集中仅考虑了32个不同的石脑油成分，所以只覆盖流出物空间的特定区域不足为奇。

5. 结论与展望

本文建立了一个由4个相互作用的深度学习人工神经网络组成的框架，根据有限数量的商业的（或可轻易获取的）石脑油特性和工艺描述符，预测石脑油性质和详细的蒸汽裂解炉流出物成分。每个单独的网络都能达到优异性能，与传统在线分析设备和COILSIM1D等商用工具的准确性不相上下，甚至超过了它们。根据石脑油的PIONA特性、密度和蒸汽压重构详细的原料组成，用两个深度学习人工神经网络，获得28个不同（伪）成分的平均绝对误差均值为0.36%。使用真实、详细石脑油组成预测流出物组成时，平均绝对误差的均值为0.13%；使用根据上述指标重构的石脑油组成预测流出物组成时，平均绝对误差的均值为0.19%。预测精度高，加上计算成本非常低（整个框架的执行以毫秒为单位），让建立的网络非常适用于工艺参数难以获取的实时监测。这些网络也适用于新的实时优化算法，过程调整频率比当前要高得多。在毫秒为单位的计算延迟下，甚至可以考虑在前馈过程控制中应用。虽然使用反应器和炉的特定配置对提出的网络进行了模拟训练，但是输入中包含与反应器无关的强度指数，使得网络本身与反应器无关。结果表明，所述方法适用于任何类型的反应器，而无任何性能损失。深度学习人工神经网络的主要缺点是丢失了问题的物理意义和可解释意义。对于工艺和工艺设计背后的复杂化学机制的详细因果分析，详细的动力学模型仍然至关重要。提出的模型已经过模拟数据训练，这一事实进一步推动了基本模型的发展。然而，在上述实时优化与工艺控制等许多实际应用中，执行速度、准确性和易用性是主要关注点。由于深度学习人工神经网络灵活且具有强大的预测能力，所以影响工厂优化的蒸汽裂解过程的其他几个方面，如积碳等，在将来也可以用类似的方法进行处理。

Acknowledgements

Pieter P. 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 financial support from the Long Term Structural Methusalem Funding by the

Flemish Government (BOF09/01M00409).

Compliance with ethics guidelines

Pieter P. Plehiers, Steffen H. Symoens, Ismaël Amghizar, Guy B. Marin, Christian V. Stevens, and Kevin M. Van Geem 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, 1bar = 10^5 Pa)
COT	coil outlet temperature (K)
CPD	cyclopentadiene
CPU	central processing unit
DL	deep learning (> 1 hidden layer)
E/E	ethylene/ethane ratio
FBP	final boiling point (K)
GC×GC	GC two-dimensional gas chromatography
GPU	graphics processing unit
IBP	initial boiling point (K)
M/P	methane/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, *iso*-paraffins, olefins, naphthenes, aromatics
ReLU	rectified linear unit
RF	random forest
(R)MSD	(root) mean square deviation
RTO	real-time optimization
SVR	support vector regression

Variables

$\boldsymbol{A}$	matrix of eigenvectors
A_k	aromatics with k carbon atoms
b	perceptron/layer bias
C_k	hydrocarbons with k carbon atoms
d	(chosen) dimensionality of the PC space
f	activation function
$F_{a,p,n}$	F-statistic with confidence level a, p degrees of freedom, and n samples
i	perceptron/layer input
$\boldsymbol{i}$	perceptron/layer input vector
I_k	*iso*-paraffins with k carbon atoms
n	number of data points in dataset
N_k	naphthenes with k carbon atoms
$\boldsymbol{o}$	layer output
o	perceptron output
O_k	olefins with k carbon atoms
P_k	paraffins with k carbon atoms
$\boldsymbol{S}$	variance-covariance matrix of the dataset
w	weight
$\boldsymbol{W}$	weight matrix for single layer
$\boldsymbol{w}$	weight vector for single perceptron
x	model input
$\boldsymbol{x}$	model input vector
y	model output
$\boldsymbol{y}$	model output vector
$\boldsymbol{z}$	input representation in the PC space
α	probability level
$\boldsymbol{\Lambda}$	diagonal matrix of eigenvalues
λ	eigenvalue
$\boldsymbol{\Lambda'}$	eigenvector matrix in the reduced-dimension PC space

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2019.02.013.

References

[1] Amghizar I, Vandewalle LA, Van Geem KM, Marin GB. New trends in olefin production. Engineering 2017;3(2):171–8.

[2] Campbell M, Hoane AJ Jr, Hsu F. Deep blue. Artif Intell 2002;134(1–2):57–83.
[3] Gibney E. Google AI algorithm masters ancient game of Go. Nature 2016;529 (7587):445–6.
[4] Chowdhury GG. Natural language processing. Annu Rev Inf Sci Technol 2003;37(1):51–89.
[5] Yin W, Kann K, Yu M, Schütze H. Comparative study of CNN and RNN for natural language processing. 2017. arXiv:1702.01923.
[6] Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, et al. End to end learning for self-driving cars. 2016. arXiv:1604.07316.
[7] Li D, Gao H. A hardware platform framework for an intelligent vehicle based on a driving brain. Engineering 2018;4(4):464–70.
[8] Maltarollo VG, Honório KM, Ferreira da Silva AB. Applications of artificial neural networks in chemical problems. In: Suzuki K, editor. Artificial neural networks—architectures and applications. Rijeka: InTech; 2013. p. 203–23.
[9] Day CP. Robotics in industry—their role in intelligent manufacturing. Engineering 2018;4(4):440–5.
[10] Brettel M, Friederichsen N, Keller M, Rosenberg M. How virtualization, decentralization and network building change the manufacturing landscape: an Industry 4.0 perspective. Int J Inf Commun Eng 2014;8(1):37–44.
[11] Lasi H, Fettke P, Kemper HG, Feld T, Hoffmann M. Industry 4.0. Bus Inf Syst Eng 2014;6(4):239–42.
[12] Zhong RY, Xun X, Klotz E, Newman ST. Intelligent manufacturing in the context of Industry 4.0: a review. Engineering 2017;3(5):616–30.
[13] Zhou K, Liu T, Zhou L. Industry 4.0: towards future industrial opportunities and challenges. In: Proceeding of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD); 2015 Aug 15–17; Zhejiang, China. New York: IEEE; 2015. p. 2147–52.
[14] Yuan Z, Qin W, Zhao J. Smart manufacturing for the oil refining and petrochemical industry. Engineering 2017;3(2):179–82.
[15] Zhang L, Mao H, Liu L, Du J, Gani R. A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. Comput Chem Eng 2018;115:295–308.
[16] Bajorath J. Computer-aided drug discovery. F1000Res 2015;4:630.
[17] Peplow M. Organic synthesis: the robo-chemist. Nature 2014;512(7512):20–2.
[18] Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model 2018;58(2):252–61.
[19] Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem 2017;38(16):1291–307.
[20] Sedghi S, Huang B. Real-time assessment and diagnosis of process operating performance. Engineering 2017;3(2):214–9.
[21] Bogle IDL. A perspective on smart process manufacturing research challenges for process systems engineers. Engineering 2017;3(2):161–5.
[22] Castillo PAC, Castro PM, Mahalec V. Global optimization of nonlinear blendscheduling problems. Engineering 2017;3(2):188–201.
[23] Van Geem KM, Pyl SP, Reyniers MF, Vercammen J, Beens J, Marin GB. On-line analysis of complex hydrocarbon mixtures using comprehensive twodimensional gas chromatography. J Chromatogr A 2010;1217(43):6623–33.
[24] Van Geem KM, Marin G, Muñoz Gandarillas A, Zhang Y, Du W, Qian F. Plant wide optimization for high value added products: a steam cracking case study [presentation]. In: The 30th Ethylene Producers' Conference; 2018 Apr 22–26; Orlando, FL, USA; 2018.
[25] Hudebine D, Verstraete JJ. Molecular reconstruction of LCO gasoils from overall petroleum analyses. Chem Eng Sci 2004;59(22–23):4755–63.
[26] Verstraete JJ, Revellin N, Dulot H, Hudebine D. Molecular reconstruction of vacuum gasoils. ACS Div Fuel Chem 2004;49(1):20–1.
[27] 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(9):1020–34.
[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. Prog Energy Combust Sci 2001;27 (1):99–139.
[29] Sadrameli SM. Thermal/catalytic cracking of hydrocarbons for the production of olefins: a state-of-the-art review I: thermal cracking review. Fuel 2015;140:102–15.
[30] Van Geem KM, Reyniers MF, Marin GB. Challenges of modeling steam cracking of heavy feedstocks. Oil Gas Sci Technol 2008;63(1):79–94.
[31] Van Geem KM, Reyniers MF, Marin GB. Two severity indices for scale-up of steam cracking coils. Ind Eng Chem Res 2005;44(10):3402–11.
[32] Van Geem KM, Žajdlík R, Reyniers MF, Marin GB. Dimensional analysis for scaling up and down steam cracking coils. Chem Eng J 2007;134(1–3):3–10.
[33] Van Geem KM, Reyniers MF, Pyl S, Marin GB, Zhou Z. Effect of operating conditions and feedstock composition on run lengths of steam cracking coils [presentation]. In: AIChE Spring National Meeting; 2009 Apr 26–30; Tampa, FL, USA; 2009.
[34] Green WH Jr. Predictive kinetics: a new approach for the 21st century. Adv Chem Eng 2007;32:1–50.
[35] Van de Vijver R, Vandewiele NM, Bhoorasingh PL, Slakman BL, Seyedzadeh Khanshan F, Carstensen HH, et al. Automatic mechanism and kinetic model generation for gas- and solution-phase processes: a perspective on best practices, recent advances, and future challenges. Int J Chem Kinet 2015;47 (4):199–231.
[36] Hopfield JJ. Artificial neural networks. IEEE Circuits Device 1988;4(5):3–10.
[37] Mahanta J. Introduction to neural networks, advantages and applications [Internet]. Deeplearningtrack; [updated 2017 Jul 9; cited 2018 Aug 3]. Available form: https://www.deeplearningtrack.com/single-post/2017/07/09/Introduction-to-NEURAL-NETWORKS-Advantages-and-Applications.
[38] Pyl SP, Van Geem KM, Reyniers MF, Marin GB. Molecular reconstruction of complex hydrocarbon mixtures: an application of principal component analysis. AIChE J 2010;56(12):3174–88.
[39] Niaei A, Towfighi J, Khataee AR, Rostamizadeh K. The use of ANN and the mathematical model for prediction of the main product yields in the thermal cracking of naphtha. Pet Sci Technol 2007;25(8):967–82.
[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. Ind Eng Chem Res 2011;50(3):1536–47.
[41] Ghadrdan M, Mehdizadeh H, Boozarjomehry RB, Darian JT. On the introduction of a qualitative variable to the neural network for reactor modeling: feed type. Ind Eng Chem Res 2009;48(8):3820–4.
[42] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems; 2013 Dec 5–8; Lake Tahoe, NV, USA. San Diego: NIPS; 2013. p. 2553–61.
[43] Seif G. I'll tell you why Deep Learning is popular in demand [Internet]. Medium; [cited 2018 Aug 3]. Available from: https://medium.com/swlh/illtell- you-why-deep-learning-is-so-popular-and-in-demand-5aca72628780.
[44] Shamsuddin SM, Ibrahim AO, Ramadhena C. Weight changes for learning mechanisms in two-term back-propagation network. In: Suzuki K, editor. Artificial neural networks—architectures and applications. Rijeka: InTech; 2013. p. 53–82.
[45] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by backpropagating errors. Nature 1986;323(6088):533–6.
[46] Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci 1995;35(5):826–33.
[47] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw 1989;2(5):359–66.
[48] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA. San Diego: NIPS; 2012. p. 1097–105.
[49] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.
[50] Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the 21th International Conference on Machine Learning; 2004 Jul 4–8; Banff, AB, Canada. New York: ACM; 2004. p. 78.
[51] Chollet F. Keras: the Python deep learning library [Internet]. [cited 2018 Aug 3]. Available from: https://keras.io.
[52] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation; 2016 Nov 2–4; Savannah, GA, USA. Berkeley: USENIX Association; 2016. p. 265–83.
[53] Jolliffe IT. Principal component analysis. New York: Springer; 2011.
[54] De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst 2000;50(1):1–18.
[55] Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
[56] Van Geem KM, Reyniers MF, Marin G. Taking optimal advantage of feedstock flexibility with COILSIM1D. In: Proceedings of 2008 AIChE Spring Meeting and Global Congress on Process Safety; 2008 Apr 6–10; New Orleans, LA, USA. New York: American Institute of Chemical Engineers; 2008. p. 391–404.
[57] Vervust A, Amghizar I, Munoz A, Van Geem KM, Marin G. Full furnace simulations and optimization with COILSIM1D. In: Proceedings of 2016 Spring Meeting and 12th Global Congress on Process Safety; 2016 Apr 10–14; Houston, TX, USA. New York: American Institute of Chemical Engineers; 2016. p. 21.
[58] Paraskevas PD, Sabbe MK, Reyniers MF, Marin GB, Papayannakos NG. Group additive kinetic modeling for carbon-centered radical addition to oxygenates and b-scission of oxygenates. AIChE J 2016;62(3):802–14.
[59] Saeys M, Reyniers MF, Marin GB, Van Speybroeck V, Waroquier M. Ab initio group contribution method for activation energies for radical additions. AIChE J 2004;50(2):426–44.
[60] Van de Vijver R, Sabbe MK, Reyniers MF, Van Geem KM, Marin GB. Ab initio derived group additivity model for intramolecular hydrogen abstraction reactions. Phys Chem Chem Phys 2018;20(16):10877–94.
[61] Davis AC, Francisco JS. Ab initio study of hydrogen migration across n-alkyl radicals. J Phys Chem A 2011;115(14):2966–77.
[62] Gao CW, Allen JW, Green WH, West RH. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. Comput Phys Commun 2016;203:212–25.
[63] Merchant SS. Molecules to engines: combustion chemistry of alcohols and their applications to advanced engines [dissertation]. Cambridge: Massachusetts Institute of Technology; 2015.
[64] Fannin G. Distillation process analyser with ASTM 86 compliance. Petro Industry News 2013 Aug/Sep;14(4):40.
[65] Ferris AM, Rothamer DA. Methodology for the experimental measurement of vapor–liquid equilibrium distillation curves using a modified ASTM D86 setup. Fuel 2016;182:467–79.