


Scale-up of complex molecular reaction system by hybrid mechanistic modeling and deep transfer learning

Received: 2 December 2024

Zhengyu Chen, Yongqing Xie, Chunming Xu & Linzhou Zhang  

Accepted: 29 August 2025

Published online: 07 October 2025

 Check for updates

The scale-up of chemical processes involves substantial changes in reactor size, operational modes, and data characteristics, leading to significant challenges in predicting product distribution across scales. This study presents a unified modeling framework that integrates the mechanistic model with deep transfer learning to accelerate chemical process scale-up. The framework is demonstrated through a case study on naphtha fluid catalytic cracking. A molecular-level kinetic model was developed from laboratory-scale experimental data, and a deep neural network was designed and trained to represent complex molecular reaction systems. To address the challenge of discrepancies in data types at various scales, a property-informed transfer learning strategy was developed by incorporating bulk property equations into the neural network. This approach enabled automated prediction of pilot-scale product distribution with minimal data. Moreover, process conditions of the pilot plant were optimized using a multi-objective optimization algorithm.

Process scale-up is a critical step in advancing chemical processes from laboratory research to industrial production, involving step-by-step experiments at laboratory, pilot, and industrial scales. This process is highly complex due to variations in reactor type and mode of operation (batch to continuous), significantly affecting apparent reaction rates and transport phenomena. These complexities make the scale-up process time-intensive and expensive, posing a persistent challenge for the commercialization of new chemical processes. The traditional research approach, combining experiments and simulations, has been effectively applied to simple systems¹. However, for complex molecular reaction systems, such as coal, petroleum, and biomass, current simulation tools struggle to accurately model feedstock compositions and reaction mechanisms at the molecular level. This limitation hinders the direct scale-up of complex molecular reaction systems from laboratory to industrial scale.

To accurately describe the molecular conversion behavior, several modeling methods were proposed to simulate complex reaction systems at the molecular level, including the single-event kinetic model², the bond-electron matrix (BEM)^{3,4}, the structure-oriented lumping (SOL)^{5,6}, the molecular type and homologous series (MTHS) matrix⁷,

and the structural unit and bond-electron matrix (SU-BEM)⁸. The single-event kinetic model is extensively used in acid-catalyzed reaction systems, such as fluid catalytic cracking (FCC)^{9,10} and methanol-to-olefins (MTO)¹¹ processes. The BEM is applied in biomass and petroleum conversion processes^{12–14}, while SOL, MTHS, and SU-BEM frameworks are predominantly applied in petroleum refining^{15–18}. These methodologies tackle core challenges in complex reaction systems, including molecular compositional modeling¹⁹, reaction network generation^{20,21}, kinetic model development, and model parameter organization²². The reaction behavior of species at the mechanistic and pathway levels can be modeled with high precision.

High-precision molecular-level kinetic models are essential for accelerating the process scale-up of complex molecular reaction systems. The computational accuracy of the model depends primarily on the kinetic parameters, which are influenced by reactor dimensions. For example, kinetic parameters derived from a laboratory-scale reactor cannot directly predict the production distribution in a pilot or industrial plant. This challenge arises from the combined effects of the chemical reaction and the transport phenomenon on the product distribution, as variations in reactor size and structure change the

State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum, Beijing, PR China.

✉ e-mail: Lzz@cup.edu.cn

transfer rate and apparent kinetic parameters. FCC, a representative complex molecular reaction system, highlights these challenges. Changes in reactor types (from fixed fluidized bed to riser) and operating modes (from batch to continuous operation) require the development of corresponding models for each reactor scale. Currently, there is no universal cross-scale computational method to accommodate reactors of varying sizes²³. To bridge this gap, researchers have coupled one-dimensional riser reactor models with molecular-level kinetic models²⁴. These developed models can predict the molecular concentration, temperature, and pressure distribution along the axial direction of the riser. However, they fail to account for the radial gradient distribution within the riser, limiting their generalizability. Computational fluid dynamics (CFD) was employed to simulate the velocity, temperature, and yield distributions in the riser, revealing that the significant influence of reactor scale on flow regimes^{25,26}. However, the highly nonlinear nature of the flow regime in the riser poses a challenge related to computational efficiency. Moreover, CFD relies on lumped kinetic models, which lack molecular-level reaction information²⁷.

The nonlinearity and computational inefficiency of the transport phenomena models present significant challenges for cross-scale computation in complex molecular reaction systems. Artificial intelligence (AI) offers substantial advantages in addressing nonlinear problems and improving computational efficiency, and AI-mechanism hybrid models have emerged as a promising approach to address these challenges^{28–31}. Integrating mechanistic models directly into neural networks to build hybrid model has attracted broad attention, such as physics-informed neural network (PINN)³² and neural ordinary differential equation (Neural ODE)³³. These methods have also been applied in some simple reaction systems³⁴. However, extending them to more complex systems remains difficult, since incorporating a large number of ODEs into neural networks still poses significant challenges for model training. An alternative hybrid modeling strategy is data-driven models. This model is trained using data generated from mechanistic models. Hough et al.³⁵ utilized the artificial neural network (ANN) to build a data-driven model for biomass pyrolysis, facilitating a rapid solution for large-scale reaction networks. Qiu and collaborators³⁶ introduced a data-driven model for the naphtha steam cracking process based on the convolutional neural network (CNN). Compared to ANN, CNN can recognize features in the reaction network and accelerate computations by approximately 300 times.

Despite some progress made in AI-mechanism hybrid models, most applications remain limited to single-scale systems, and efficient cross-scale computational methods for process scale-up are still lacking. Recently, Chen et al.³⁷ proposed a parameter transfer strategy. The strategy utilized pilot-scale data to fine-tune the kinetic parameter regressed from a laboratory-scale reactor, achieving cross-scale computation for the FCC process. This work provides new insights for cross-scale modeling of complex conversion systems, and replacing the parameter transfer strategy with transfer learning may be a more efficient and generalizable solution³⁸. Transfer learning aims to transfer knowledge from a well-learned task (source domain) to a related but distinct task (target domain), enhancing model performance or reducing data requirements. It has been widely applied in chemical engineering, such as property prediction³⁹, fault diagnosis⁴⁰, and process simulation⁴¹. For example, Buterez et al.³⁹ developed a graph neural network (GNN)-based transfer learning method to predict expensive molecular properties by inexpensive, widely available property data. Xiao et al.⁴¹ proposed a recurrent neural network (RNN)-based transfer learning framework to simulate the operations of various continuous stirred-tank reactors (CSTRs). In chemical process scale-up, although apparent reaction rates vary due to changes in transport phenomena, intrinsic reaction mechanisms remain unchanged across scales. This characteristic agrees well with the principles of transfer learning. If transfer learning can effectively capture flow

regime variations across scales, it is expected to solve the cross-scale computational challenges in reaction processes.

Therefore, this work presents a method by the hybrid mechanistic model with deep transfer learning (hybrid model). The proposed methodology was applied to the naphtha FCC process from the laboratory to the pilot plant. A molecular-level kinetic model (mechanistic model) of the naphtha FCC was developed using data from a laboratory-scale reactor. Subsequently, a neural network architecture for transfer learning of complex reaction systems was designed. The neural network was trained by data generated from the mechanistic model, and a laboratory-scale data-driven model was obtained. To accurately calculate the pilot-scale product bulk properties, mechanistic equations for calculating bulk properties were incorporated into the neural network. Network parameters were fine-tuned using limited pilot plant data. The developed model provides a foundational computational tool for the scale-up of complex reaction processes.

Result

Hybrid mechanistic modeling and deep transfer learning for process scale-up

The primary reason for changes in product distribution during the scale-up process is the variation in apparent reaction rates, caused by changes in transport phenomenon, but the intrinsic reaction mechanisms remain almost unchanged. According to this characteristic, we propose describing the reaction mechanism by an easy-to-model kinetic model. For the hard-to-model transport phenomenon, we use transfer learning to automatically capture the changing features. On this basis, A hybrid model integrated mechanistic model with transfer learning is developed, as illustrated in Fig. 1.

The methodology begins with the development of a molecular-level kinetic model, built using detailed product distribution data obtained under various laboratory conditions. A series of molecular conversion datasets with varying compositions and conditions was generated using the molecular-level kinetic model. Generated data were used to train the neural network, and the laboratory-scale data-driven model was built. To adapt the model for pilot- and industrial-scale applications, transfer learning was employed. The pilot- or industrial-scale datasets were expanded through data augmentation. Partially hidden layers in the neural network were fine-tuned with the augmented pilot or industrial data to accurately calculate the product distribution, achieving cross-scale computation for complex molecular reaction systems.

The proposed hybrid model offers a significant perspective on process scale-up. However, applying this method to complex molecular reaction systems still presents several challenges:

- (1) Optimizing network architectures for transfer learning: Existing data-driven models are typically built using a single neural network. Transfer learning often relies on a trial-and-error approach to determine which parameters to freeze or fine-tune, and the process mechanism offers limited guidance for parameter fine-tuning. It is necessary to develop transfer learning network architectures suitable for complex molecular reaction systems.
- (2) Data discrepancies at different scales: Laboratory-scale data (source domain) often includes detailed molecular-level characterization information, while pilot and industrial plants (target domain) primarily provide product bulk properties. This mismatch poses challenges for applying transfer learning at different scales.

To overcome these limitations, this study proposed a deep transfer learning network architecture suitable for complex molecular reaction systems. Additionally, we developed a property-informed transfer learning strategy by integrating the bulk property equations

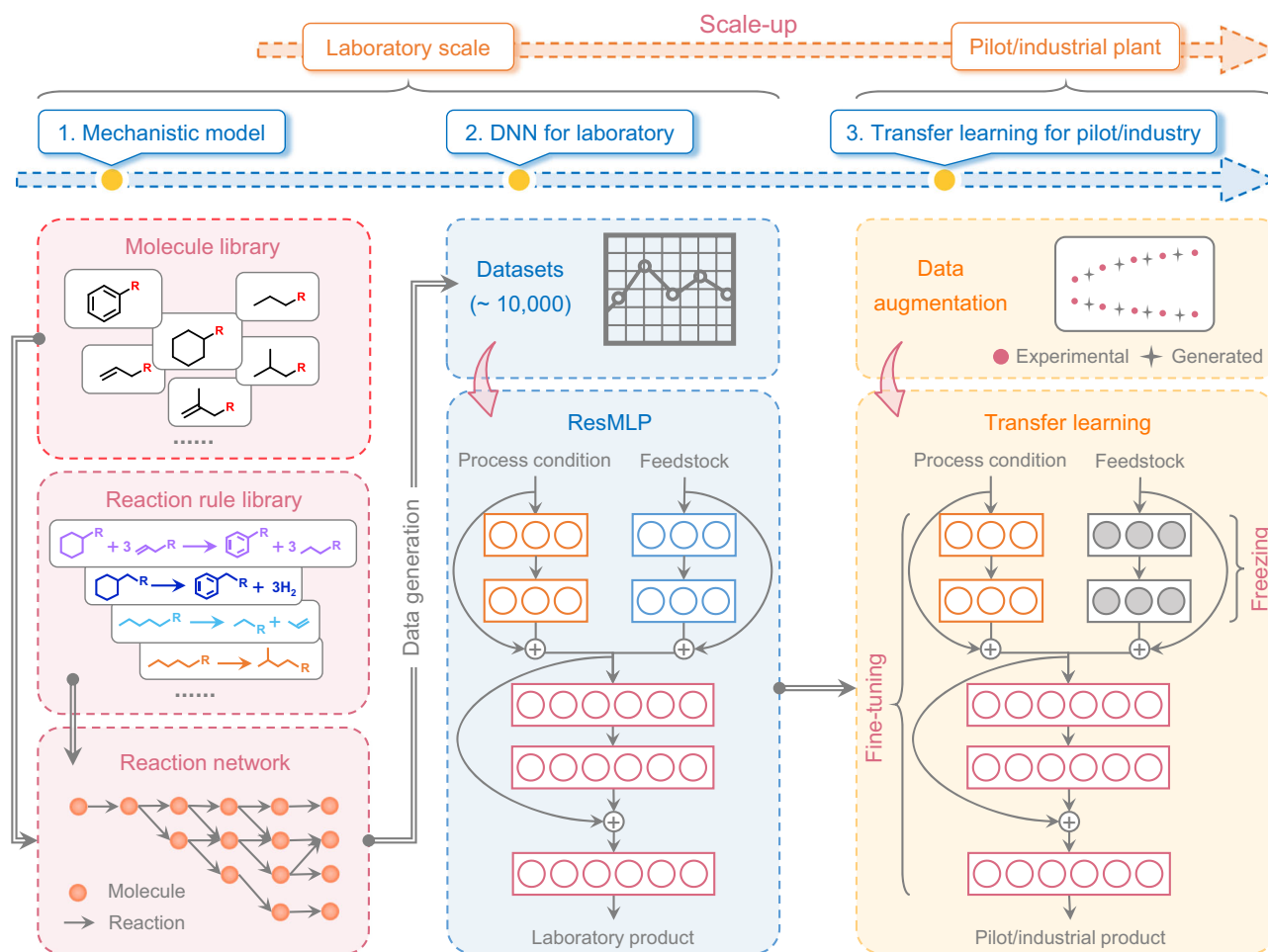


Fig. 1 | A modeling framework for hybrid mechanistic modeling and deep transfer learning. This figure outlines the process for developing a hybrid model. Initially, a molecular-level kinetic model for naphtha FCC is developed as the mechanistic model. The model is then used to generate molecular conversion data under various process conditions and feedstock types, which are employed to train

a residual multi-layer perceptron (ResMLP) suitable for process scale-up of complex reaction systems. On this basis, small-sample pilot or industrial data are used to fine-tune partial neural network parameters, enabling accurate prediction of product distribution for a pilot or industrial-scale plant.

into the neural network to bridge the data gap between laboratory and larger-scale operations.

Network architecture of transfer learning for complex molecular reaction systems

To address the challenge that a priori knowledge cannot effectively guide the parameter fine-tuning process, this work introduces a deep transfer learning network architecture suitable for complex reaction processes, as shown in Fig. 2a. The proposed architecture simulates the computational procedure of the mechanistic model by integrating three residual multi-layer perceptron (ResMLPs). In a mechanistic model, process conditions and feedstock composition serve as inputs to calculate the product distribution, respectively. Similarly, the first ResMLP (Process-based ResMLP) is used to input process conditions. The second network (Molecule-based ResMLP) takes molecular composition data as input to capture compositional features. The outputs from these two networks are then combined in an Integrated ResMLP to predict the product molecular composition. This architecture mirrors the logic of the mechanistic model, facilitating more precise identification of parameter fine-tuning locations during transfer learning. For example, when the feedstock composition and catalyst remain unchanged but the reactor structure and process conditions are adjusted, the Molecule-based ResMLP can be frozen, and only the Process-based and Integrated ResMLPs require fine-tuning.

Conversely, if one intends to investigate the product distribution of different feedstocks in the same reactor, the Process-based ResMLP should remain frozen, while only the Molecule-based and Integrated ResMLPs need fine-tuning.

Property-informed transfer learning strategy for process scale-up

In addition to parameter fine-tuning, the mismatch between laboratory-scale molecular composition data and pilot/industrial bulk property data also presents a challenge for direct transfer learning. Although modifying the output layer of the ResMLP to predict bulk properties is an option, this approach inevitably loses product information at the molecular level. To accurately compute the bulk properties while retaining the molecular composition information, this work proposed a property-informed transfer learning strategy, as shown in Fig. 2b. Inspired by the PINN, we also integrated the mechanistic model (bulk property calculation equations) into the neural network. However, compared to PINN, which adds an additional regularization term to the loss function, this method directly modified the loss function to compute bulk properties. It is a post-processing step for the output layer results in the neural network training process. When the molecular composition is predicted by the output layer, the corresponding bulk properties (e.g., fraction yield and product composition) are simultaneously computed within the loss function. The

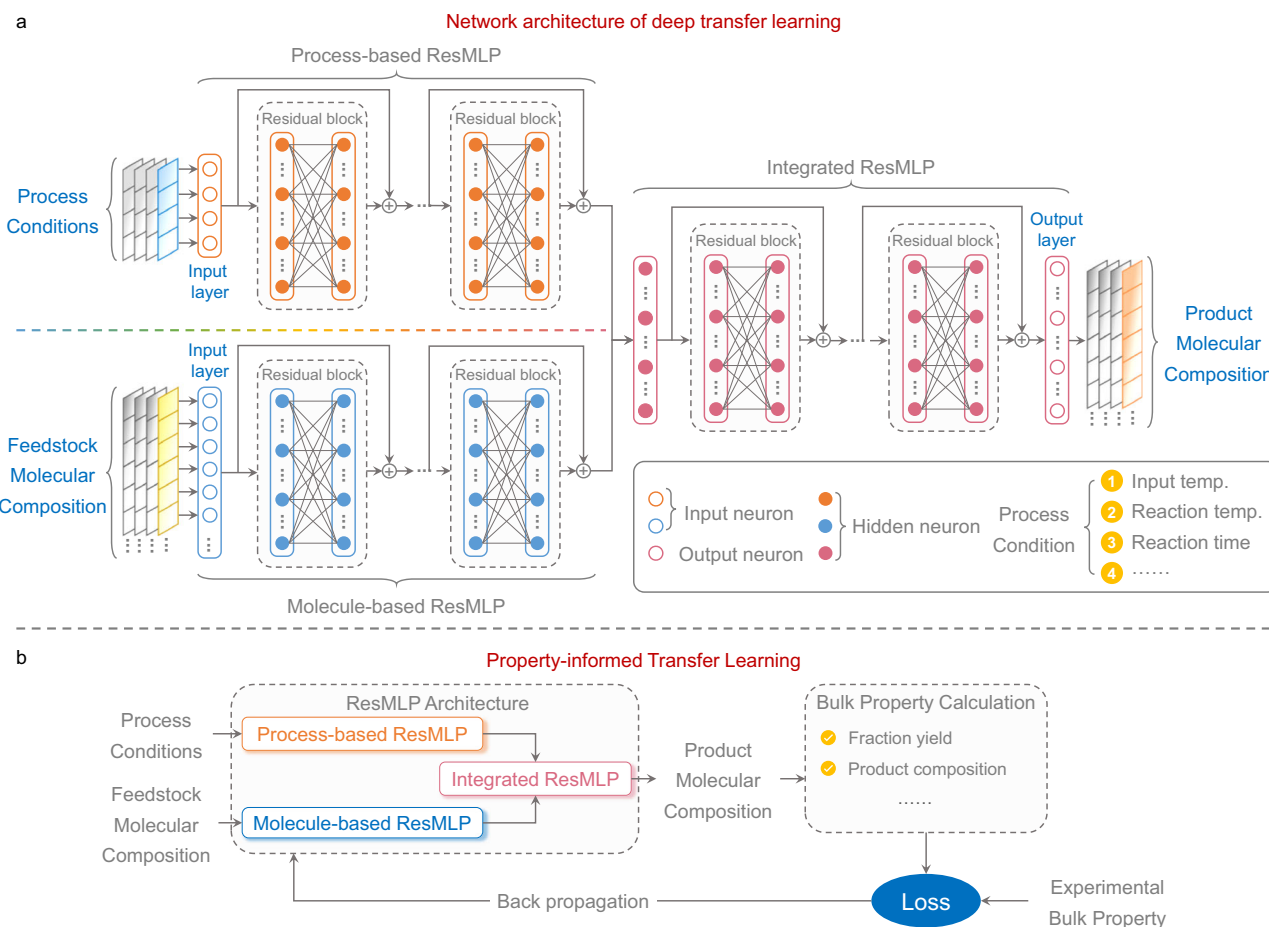


Fig. 2 | Hybrid model architecture for process scale-up of complex molecular reaction systems. a Network architecture of deep transfer learning. The neuron network takes process conditions and feedstock molecular composition as inputs and outputs the product molecular composition. The Process-based ResMLP and Molecule-based ResMLP are residual multi-layer perceptrons (ResMLPs) designed to capture features of process conditions and feedstock composition, respectively. The Integrated ResMLP combines the outputs of both networks to compute the product molecular composition. Hyperparameters within the network architecture

are optimized using Bayesian optimization algorithms. **b** Property-informed transfer learning strategy. This strategy integrates the mechanistic model for calculating bulk properties into the fine-tuning process. After the final layer of the network generates the molecular composition, bulk properties such as fraction yield and gasoline composition are estimated by bulk property calculator. Measured bulk properties are compared with the calculated values to compute the mean absolute error (MAE). Based on this error, backpropagation is used to fine-tune the neural network parameters.

experimental bulk property data are then compared to the calculated values, and the error is used to update network parameters through backpropagation. This method addresses the mismatch between data at different scales, and the model can calculate the product bulk properties and predict the molecular composition.

Molecular-level kinetic model for naphtha FCC

To validate the proposed hybrid model, we selected the naphtha FCC as a case study. Naphtha FCC is a typical complex molecular reaction system involving hundreds of molecules. Naphtha derived from the FCC unit contains a large number of olefins. The process aims to convert these olefins into i-paraffins, aromatics, and light olefins through hydrogen transfer and cracking reactions. The product is a high-quality gasoline blending component enriched with i-paraffins and aromatics.

Different reactors are employed for the FCC process at the laboratory and pilot/industrial plant. In the laboratory, a fixed fluidized bed (FFB) reactor is typically used, as illustrated in Fig. 3a, while pilot and industrial plants utilize a riser reactor, depicted in Fig. 3b. Although both FFB and riser are gas-solid catalytic reactions, there are significant differences in their operation modes and fluid flow regimes. The laboratory-scale reactor operates in a batch mode under isothermal conditions. Particle velocities are relatively low, and gaseous

products do not carry catalyst particles out of the reactor. In contrast, pilot and industrial reactors operate continuously under adiabatic conditions, with relatively high particle velocities. Catalyst particles are carried out of the reactor along with the gaseous products and separated in the disengager. Overall, in addition to the reaction mechanism, there are significant differences between the laboratory-scale and pilot/industrial-scale FCC units, as illustrated in Fig. 3. Directly calculating the product distribution in pilot or industrial plants using models based on laboratory-scale conversion laws is a challenging task. It is necessary to perform cross-scale computation by the hybrid model developed in this work to predict pilot-scale product distribution.

In our previous work³⁷, a laboratory-scale molecular-level kinetic model was developed based on the SU-BEM framework. The developed model can accurately predict the product molecular distribution, and the modeling approach is shown in Fig. 4a. Initially, the detailed molecular composition of naphtha 1 was characterized by gas chromatography (GC). The obtained molecular-level characterization information (Supplementary Table 1 for naphtha 1) was then transformed into a molecular compositional model⁴². According to the carbenium ion reaction mechanism, reactions in the FCC were categorized and programmed into 21 reaction rules, including cracking, isomerization, cyclization, dehydrogenation, alkylation,

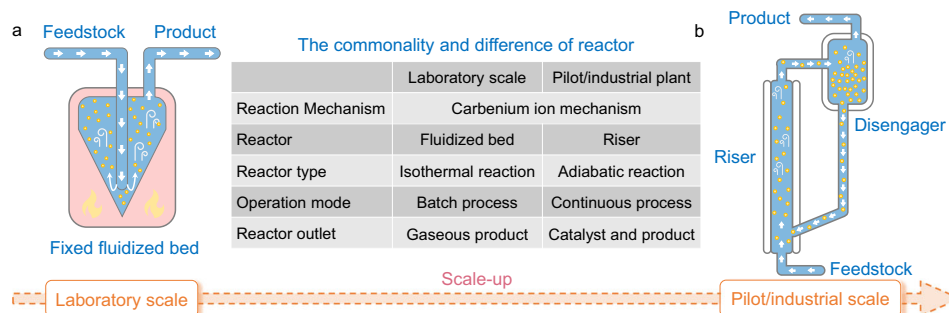


Fig. 3 | Characterization of FCC reactors at different scales. **a** Laboratory-scale FCC reactor. The reactor is an isothermal fixed fluidized bed, operating in batch mode. Gaseous products are collected at the reactor outlet. **b** Pilot/industrial-scale FCC process. The process takes place in an adiabatic riser reactor, which operates continuously. The gaseous products and deactivated catalysts are obtained at the

riser outlet and separated in a disengager. The deactivated catalyst is regenerated through coke combustion. Both laboratory-scale and pilot/industrial-scale reactors follow the carbenium ion mechanism, with the reaction mechanism remaining consistent during scale-up.

and hydrogen transfer, as listed in Supplementary Table 2. The feedstock molecules were input into the reaction rules. The reaction network auto-generated algorithm was invoked, and a complex FCC reaction network with more than 100 molecules and 750 reactions was generated, as depicted in Fig. 4a. Furthermore, the cracking reaction pathway of C8 olefins was extracted from the complete reaction network. As can be seen from the figure, olefin cyclization and hydrogen transfer were preferred for generating i-paraffins and aromatics, while the cracking reaction reduced the gasoline yield. Therefore, naphtha FCC should be carried out under milder process conditions to reduce the cracking reaction rate. Detailed information on species and reactions in the reaction network is provided in Supplementary Data 1 and 2.

According to the molecular conversion relationships, the reaction network was transformed into reaction rate equations for each substance. The reaction rates of each reaction can be calculated using the Langmuir–Hinshelwood–Hougen–Watson (LHHW) equation, and a mass balance equation coupled to catalyst deactivation was also developed. Kinetic parameters were organized via linear free energy relationship (LFER) and quantitative structure-reactivity correlation (QSRC)^{43–45}. The parameter organization method and catalyst deactivation model were discussed in Supplementary Notes 3 and 4. Experimental data under different process conditions were used to tune model parameters, and optimal kinetic parameters are listed in Supplementary Table 3. The calculated product carbon number distribution is presented in Fig. 4a. Comparing the molecular composition of feedstock and products, it was found that the FCC process converts olefins into i-paraffins and aromatics, and the calculated product conversion behavior was in agreement with the published work^{46,47}.

Hybrid models generally use process conditions and feedstock composition as inputs for neural network training. In this work, 10,000 sets of different process conditions and naphtha molecular compositions were randomly sampled. The process conditions included reaction temperature, weight hourly space velocity (WHSV), and catalyst-to-oil ratio (CTO). The data distribution in the dataset is presented in Fig. 4b. The reaction temperature primarily ranges between 400 °C and 460 °C, with a median of 430 °C. The WHSV varies from 15 to 30, and the CTO typically falls between 5 and 10, with medians of 25 and 7, respectively. These process conditions exhibit a normal distribution, encompassing typical operating ranges for naphtha FCC. Figure 4b also illustrates the naphtha composition calculated from the feedstock molecular composition. The olefin content in gasoline is the highest, predominantly distributed between 35 wt% and 39 wt%, with a median of 37 wt%. i-Paraffins and aromatics also have relatively high content, with i-paraffins ranging from 27 wt% to 30 wt%, and aromatics varying from 21 wt% to 24 wt%. The generated gasoline composition effectively

represents the typical composition of FCC naphtha. Generated data were fed into the mechanistic model to calculate the product distribution data.

Laboratory-scale hybrid model for naphtha FCC

The input-output data obtained from the molecular-level kinetic model were used to train the neural network, and the hyperparameters were optimized using a Bayesian optimization (BO) algorithm. Various evaluation metrics were compared as objective functions for hyperparameter optimization (Supplementary Table 4). The optimal hyperparameters were selected when R^2 was used as the objective function, and the ResMLP architecture based on the optimal hyperparameters is depicted in Supplementary Fig. 2. A comparison of calculated results between the mechanistic and hybrid models is presented in Fig. 5. Figure 5a, d illustrate molecular compositions predicted by the hybrid model for both the training and testing sets. The mean absolute error (MAE) of the molecular molar content was 7.23×10^{-5} and 7.75×10^{-5} , respectively. The calculated FCC product yield and gasoline composition by product molecular composition are shown in Fig. 5b, c, e, f, exhibiting excellent agreement with the mechanistic model. To validate the accuracy of the proposed network structure, we further compared the training results of different network structures, and the network structures proposed by this work can obtain better training performance (Supplementary Table 5).

After developing the hybrid model, we evaluated the effect of process conditions on product distribution, as shown in Fig. 6. The results were presented across four dimensions: fraction, component, carbon number, and molecular distribution. Overall, the hybrid model effectively captured the trends in product distribution under varying process conditions. The effect of reaction temperature on fraction yield and gasoline composition is displayed in Fig. 6a, b. As the reaction temperature increased, the cracking and condensation reaction rates rose. FCC naphtha was continuously converted to gas and heavier fractions, leading to a reduction in gasoline yield. Nevertheless, the olefin content of gasoline also gradually reduced, and the content of aromatics and i-paraffins rose. Furthermore, the influence of WHSV and CTO on product distribution was also explored (Supplementary Fig. 3).

To investigate the molecular conversion behavior, this study used the hybrid model to predict the product carbon number distribution under different reaction temperatures. Simultaneously, the calculated data from the mechanistic and hybrid models were compared with the experimental values to validate the predictive performance of the hybrid model, as shown in Fig. 6c–e. The results indicate that the predicted carbon number distribution from the hybrid model has a good agreement with the experimental measurement. i-Paraffins and

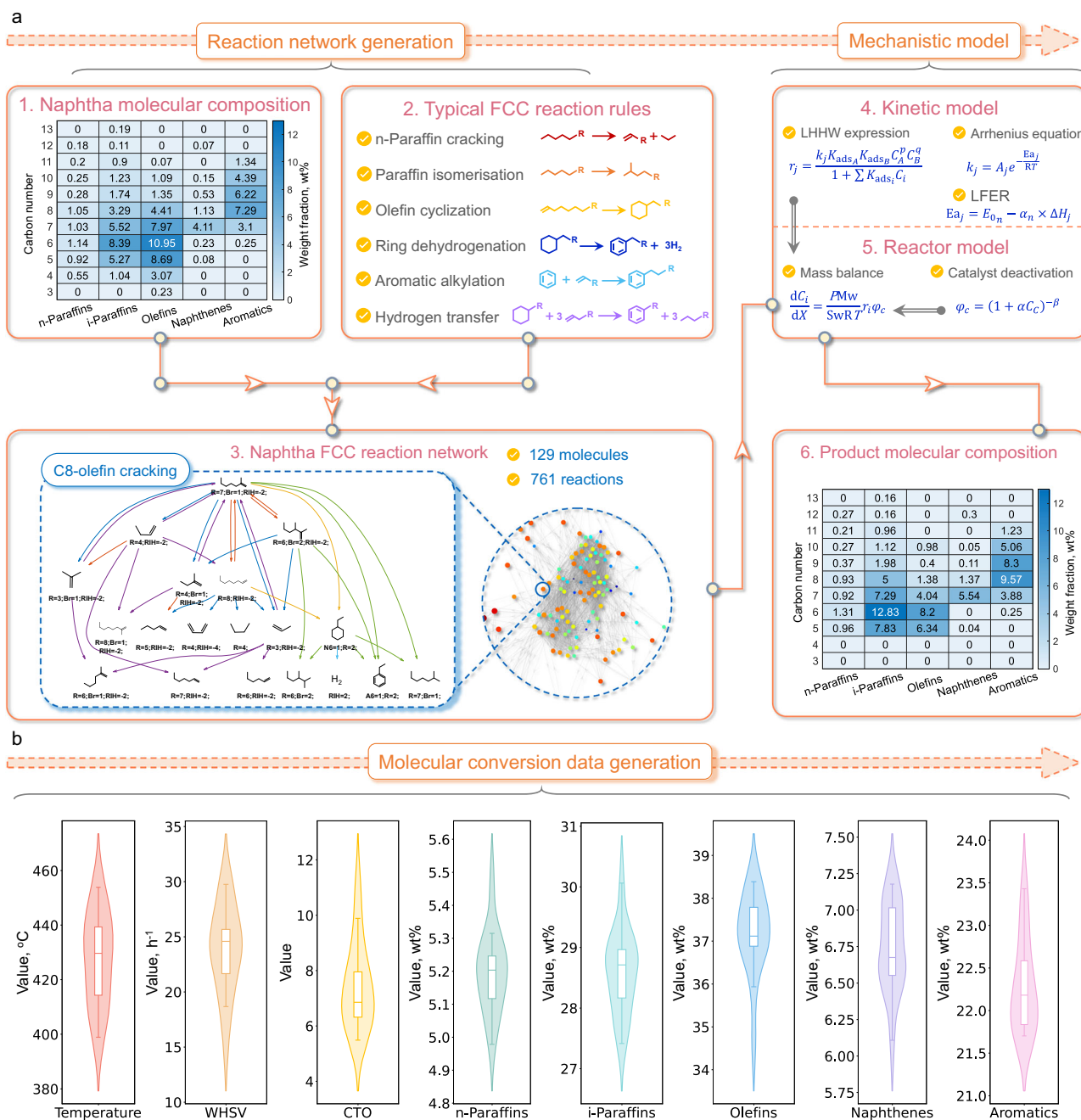


Fig. 4 | Naphtha FCC mechanistic model and data generation. **a** Modeling framework for the molecular-level kinetic model. The mechanistic model consists of two main components: reaction network generation and kinetic model development. The naphtha molecular composition matrix is input into the reaction rules derived from the carbenium ion mechanism, enabling the automatic generation of a molecular-level reaction network for naphtha FCC. This generated reaction network is then converted into a kinetic model using the Langmuir–Hinshelwood–Hougen–Watson (LHHW) formalism. The kinetic parameters are organized through the linear free energy relationship (LFER), and the kinetic model is coupled with the reactor model to

predict the molecular composition of the products. **b** Violin chart for generated molecular conversion data. The molecular conversion data consists of 10,000 samples. The left three figures display the generated process condition data, while the right five figures show the bulk properties calculated from the molecular composition. Random sampling is applied using a normal distribution function with upper and lower limits, and sampling was performed more than three times. Statistical data for the box plots, such as minima, maxima, and center, are summarized in Supplementary Table 15. Source data are provided as a Source data file.

aromatics were the dominant components, with i-paraffins primarily ranging from C5 to C7 and aromatics from C7 to C10. After the naphtha FCC, olefins with higher carbon numbers were converted, leaving predominantly C5 and C6 olefins in the product. Moreover, the molecular distributions calculated by the mechanistic model and the hybrid model were also compared to further validate the robustness of the hybrid model.

Deep transfer learning from laboratory-scale to pilot-scale
The developed laboratory-scale hybrid model successfully captured the molecular conversion behavior of naphtha FCC. Next, this work attempted to calculate the product distribution under the pilot plant by the property-informed transfer learning strategy. Taking naphtha 2 as feedstock, 15 sets of FCC pilot experiments were carried out, and product yields and gasoline composition under different process

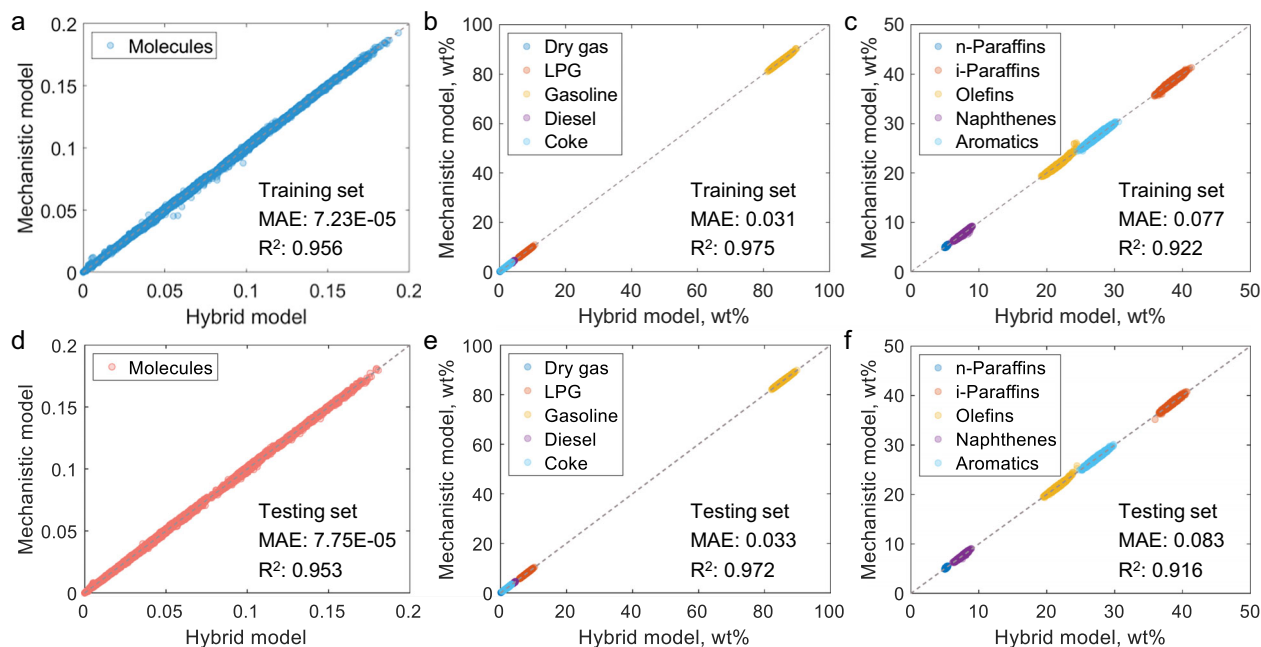


Fig. 5 | Parity plot for comparing calculated data between the hybrid model and mechanistic model. a Molecular composition in the training set. **b** Product yield in the training set. **c** Gasoline composition in the training set; **d** Molecular composition in the testing set. **e** Product yield in the testing set. **f** Gasoline composition in

the testing set. The figure also shows the mean absolute error (MAE) and R^2 values for the training, validation, and testing sets. Source data are provided as a Source data file.

conditions were obtained. Although transfer learning is effective for small-sample learning, the dataset with 15 samples remains insufficient for model training. Thus, data augmentation was performed through interpolation to expand the pilot-scale dataset. Moreover, to obtain the minimal dataset size, this work investigated the effect of dataset size on model accuracy by generating datasets with varying sample sizes: 980, 750, 500, 250, 100, 80, 60, 50, and 40 samples.

The datasets described above were used to fine-tune the network parameters. The training strategy used in this work is shown in Fig. 7(a). The complete dataset was divided into three subsets: the training set, validation set, and testing set. The generated data was divided into the training and validation sets, while experimental data were the testing set and not utilized to fine-tune the hybrid model. After completing the model training, the model was directly applied to predict the product distributions under the 15 experimental conditions, validating the robustness of the model. Since the generated laboratory-scale dataset has covered the composition distribution of naphtha 2, the hidden layers of the Molecule-based ResMLP were frozen, while the Process-based ResMLP and Integrated ResMLP were fine-tuned during the transfer learning process. Similar to the laboratory-scale hybrid model, the hyperparameters were optimized using the BO algorithm. The optimal hyperparameters for each dataset are provided in Supplementary Table 6.

The fine-tuning results for different sizes of datasets are shown in Fig. 7b, c. The blue line in the figure indicates the dataset size. When the dataset size was between 100 and 980, the MAE and mean square error (MSE) of the training, validation, and testing sets presented minimal variation. However, when the dataset size dropped below 100, both MAE and MSE gradually increased with further reductions in dataset size. This finding reveals a trade-off between dataset size and prediction accuracy. In this work, a dataset comprising 60 samples was selected as the optimal dataset, and it provides practical guidance for determining the reasonable dataset size in industrial practice.

The detailed fine-tuning results with the 60-sample dataset are illustrated in Fig. 8. The computational value of the training and

validation sets agreed well with the generated data, and the experimental product yield and gasoline composition were also accurately predicted by the fine-tuned model. The average error of bulk properties was 0.3 wt%. To further evaluate the generalization of the model, we conducted an additional pilot-scale FCC experiment using naphtha 3 as feedstock. The process conditions were as follows: feedstock temperature of 200 °C, regenerated catalyst temperature of 623 °C, reaction time of 3.6 s, and CTO of 5.5. Notably, the molecular composition of naphtha 3 differed significantly from that of naphtha 2, with a considerably higher olefin content (Supplementary Table 1). However, the developed hybrid model can still accurately predict product distribution, as shown in Fig. 8d, h. The results validated the generalization ability of the proposed model. In addition to pilot-scale product distribution prediction, the effect of the impurities and temporal product distribution was discussed in Supplementary Notes 4 and Fig. 4.

Multi-objective optimization via NN-NSGA

After the pilot-scale hybrid model was tuned, the effect of catalyst temperature on the product distribution of naphtha 2 FCC was investigated, as shown in Fig. 9a, b. With an increase in regenerated catalyst temperature, the cracking reaction rate accelerated, resulting in a reduction in gasoline yield. Concurrently, the olefins in the gasoline were progressively converted to i-paraffins and aromatics. The purpose of naphtha FCC is to reduce the olefins in gasoline while enhancing the i-paraffin yield. However, this increase in i-paraffins was accompanied by a decrease in gasoline yield. There was a trade-off between gasoline and i-paraffin yield.

To obtain the optimal process condition, a multi-objective optimization algorithm was developed by integrating the NSGA-II (non-dominated sorting genetic algorithm II) with the neural network, named NN-NSGA. Taking naphtha 2 FCC as a case study, gasoline and i-paraffin yields were defined as the optimization objectives. The Pareto frontier obtained by the NN-NSGA algorithm is shown in Fig. 9c. Among the Pareto-optimal solutions, the process condition with

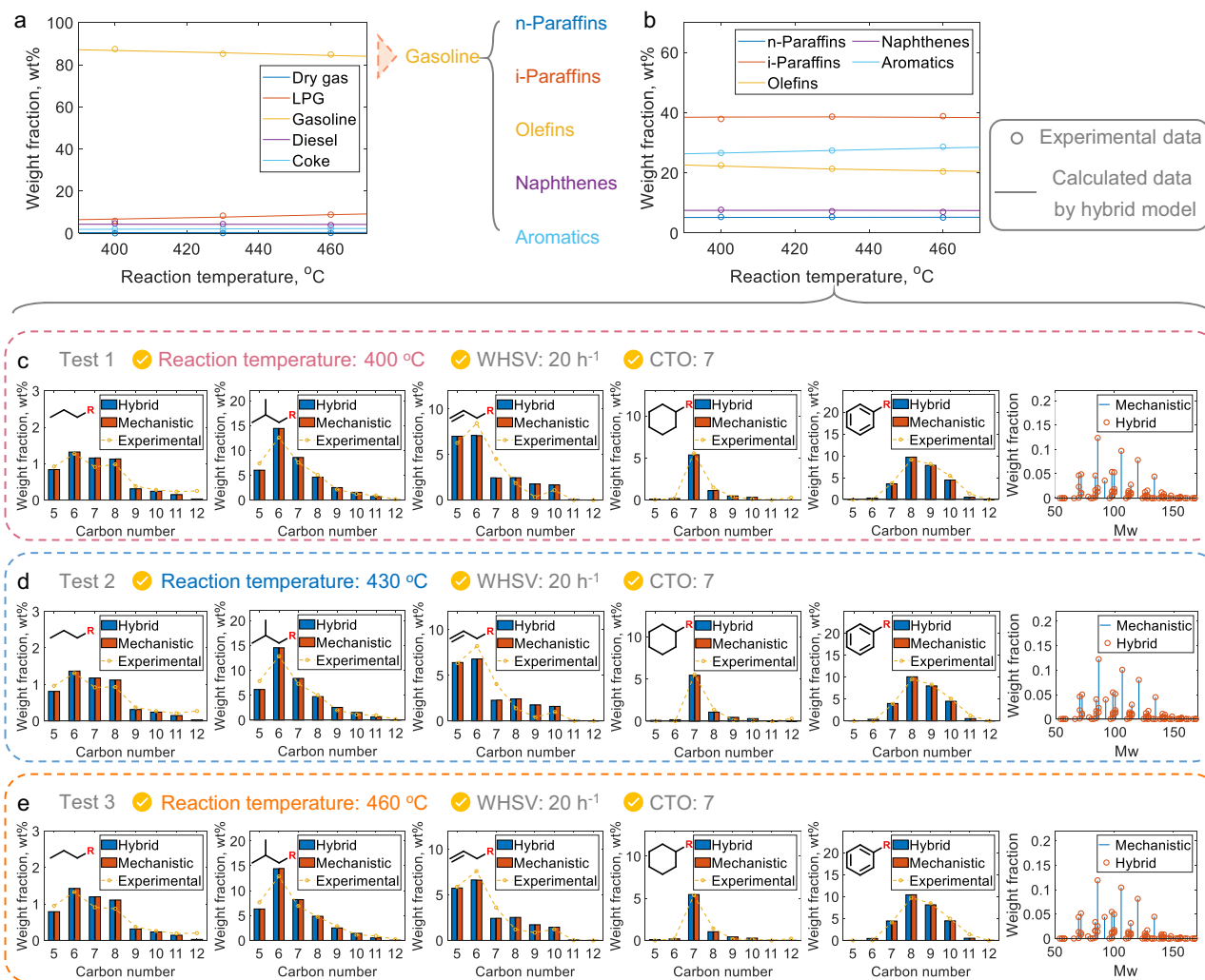


Fig. 6 | Evaluated results of the laboratory-scale hybrid model under different reaction temperatures. **a** Comparison of calculated fraction yields between the hybrid model and experimental data. Process conditions: Weight hourly space velocity (WHSV) = 20 h⁻¹, Catalyst-to-oil ratio (CTO) = 7. **b** Comparison of calculated gasoline composition between hybrid model and experimental data. Process conditions: WHSV = 20 h⁻¹, CTO = 7. **c–e** Comparison of experimental and calculated carbon number distribution and molecular composition at various reaction

temperature: 400 °C, 430 °C, and 460 °C. Process conditions: WHSV = 20 h⁻¹, CTO = 7. The left five figures compare the calculated results for n-paraffins, i-paraffins, olefins, naphthenes, and aromatics from the mechanistic and hybrid models with experimental data. The figure on the right compares the molecular composition calculated by the mechanistic and hybrid models. Source data are provided as a Source data file.

maximized i-paraffin yield was selected and compared with the experimental conditions, as shown in Table 1. The results indicate that increasing the feedstock temperature by 8 °C and reducing the catalyst temperature by 8 °C can lead to higher yields of gasoline and i-paraffins. Simultaneously, olefin and coke contents were decreased. The temperature control structure for practical implementation was proposed, as shown in Supplementary Fig. 5.

This work further compared the molecular compositions of feedstock and product under the optimal process conditions, as shown in Fig. 9d, e. The feedstock (naphtha 2) exhibited a relatively high olefin content, primarily distributed within the C5–C8 range. After the FCC process, a large number of i-paraffins and aromatics were generated from olefins via hydrogen transfer reactions. Among the products, i-paraffins with C6 exhibited the highest yield. Additionally, a series of bulk properties for the gasoline fraction was also predicted, such as distillation profile, octane number, and critical properties, as shown in Table 1. To further validate the optimized results, a molecular-level kinetic model for pilot-scale naphtha FCC was developed. A detailed introduction of this mechanistic model is provided in

Supplementary Note 13. The results indicate that the fraction yield and bulk properties predicted by the hybrid model agreed well with the rigorous mechanistic model, demonstrating the accuracy of the hybrid model. The hybrid model uncertainty caused by experiments, the mechanistic model, and the AI model was also discussed (Supplementary Note 14).

Discussion

To address the cross-scale computational challenges in the process scale-up of complex molecular reaction systems, this study proposed a hybrid model integrating molecular conversion mechanisms with deep transfer learning. The method was applied to the naphtha FCC process from the laboratory to the pilot plant. A molecular-level kinetic model was initially developed based on laboratory-scale data, serving as a data generator to produce a molecular conversion dataset with a size of $\sim 10^4$. A neural network architecture suitable for transfer learning was then designed. It incorporated two ResMLP modules to extract features from process conditions and molecular composition. This architecture achieved highly accurate predictions of molecular

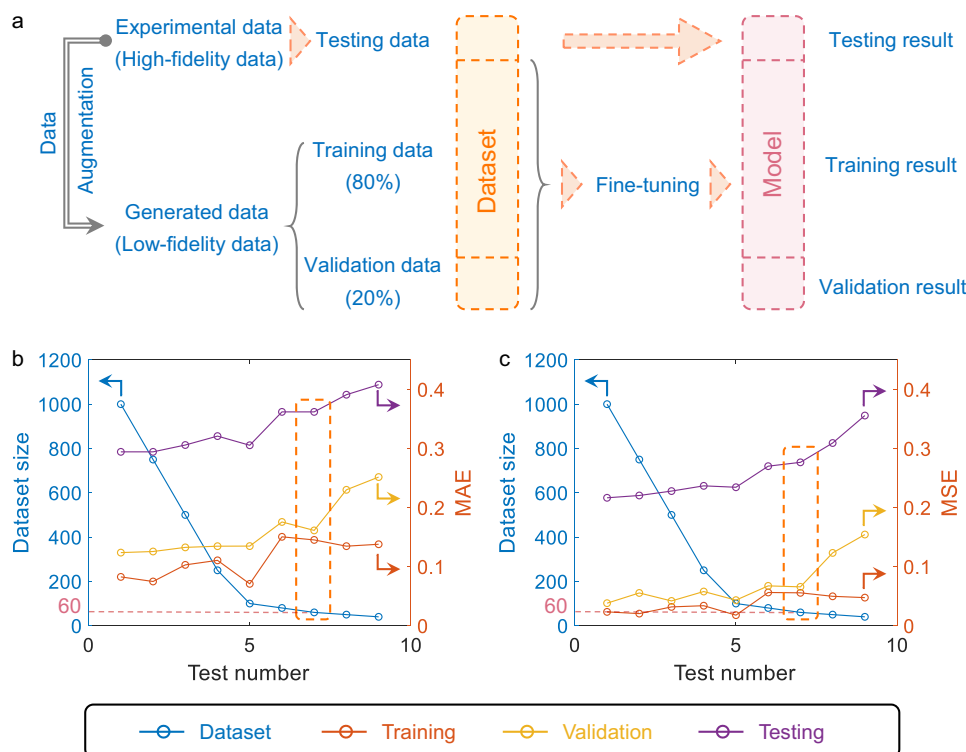


Fig. 7 | Modeling fine-tuning strategy and results for various sized datasets. **a** Modeling fine-tuning strategy. The data used for fine-tuning the model was generated from 15 sets of pilot-scale data using interpolation. The generated data was randomly divided into training and validation sets with an 8:2 ratio. After fine-tuning the model, experimental data were used as a testing set to validate the

results. **b** Mean absolute error (MAE) of training, validation, and testing sets for various sized datasets. **c** Mean square error (MSE) of training, validation, and testing sets for various sized datasets. A dataset size of 60 samples was selected as the optimal dataset for this work. Source data are provided as a Source data file.

conversion behavior within a laboratory-scale reactor, with an MAE for all molecular contents of less than 10^{-4} . To address data discrepancies at different scales, this work proposed a property-informed transfer learning strategy. The strategy can simultaneously predict product molecular composition and bulk properties using a neural network architecture, and it is beneficial to reduce time for experimental evaluation and model development/parameter regression during the FCC process scale-up. Compared to traditional molecular-level mechanistic models, the proposed method avoids modeling for each scale separately. It also improved computational accuracy over the parameter transfer strategy proposed by Chen et al.³⁷. Compared to conventional single neural network surrogate models, the proposed architecture not only improves prediction accuracy but also allows for flexible parameter freezing and fine-tuning during transfer learning.

Building on the FCC case study, we propose a generalizable modeling paradigm for process scale-up. Initially, high-throughput experiments are performed to generate data for developing the mechanistic model. The mechanistic model will be used to generate numerous data for training the hybrid model. Following this, a limited number of pilot/industrial-scale data points (as few as 4–5 points per process condition) is utilized to fine-tune the hybrid model. After fine-tuning, the model can be integrated with advanced process control systems to optimize process conditions in real-time. Notably, applying this model to other processes requires developing corresponding mechanistic models based on process characteristics. For processes dominated by transport limitations, replacing the ResMLP with PINN or Neural ODE may further enhance model accuracy, since the key transport equations can be incorporated into the neural network. Furthermore, incorporating catalyst descriptors into the hybrid model will be crucial when scale-up involves catalyst variations. Overall, this work presents a modeling framework for the scale-up of complex

reaction systems. It offers a promising pathway for complex process industries that suffer from scale-up challenges.

Methods

Laboratory-scale Experiment for Naphtha FCC

Laboratory-scale experiments of naphtha FCC were performed in an FFB reactor, which is a batch reaction device. The lower part of the reactor is conical, while the upper part is cylindrical. The reactor parameters and experimental scheme are presented in Supplementary Table 9 and Supplementary Fig. 7, respectively. One hundred and fifty grams of zeolite catalyst was initially loaded in the reactor and prefluidized with steam, and the catalyst information is listed in Supplementary Table 10. Then, the reactor was preheated to the reaction temperature, while the naphtha feedstock was pumped into the reactor bottom. The mass of naphtha feedstock was calculated by the CTO. At the reactor bottom, the feedstock contacted the catalysts and underwent the catalytic cracking reaction to generate the gaseous product. When the reaction was completed, a 20-min steam stripping process with a constant rate was implemented. The gaseous products were obtained at the condenser top via the water displacement method, and the mass can be calculated by volume. Liquid phase products were directly weighed to determine their mass. The gaseous products were divided into dry gas (C0–C2) and LPG (C3–C4) according to the carbon number, while the liquid phase was fractionated via micro-distillation into gasoline (Initial boiling point -200°C) and diesel fractions (200°C – 350°C). After cooling the reactor to ambient temperature, the coked catalyst was unloaded to determine coke yield. According to the above process, we selected naphtha 1 as feedstock and performed 8 sets of laboratory-scale FCC experiments. The composition of naphtha 1 and process conditions were presented in Supplementary Table 1 and Supplementary Table 11, respectively.

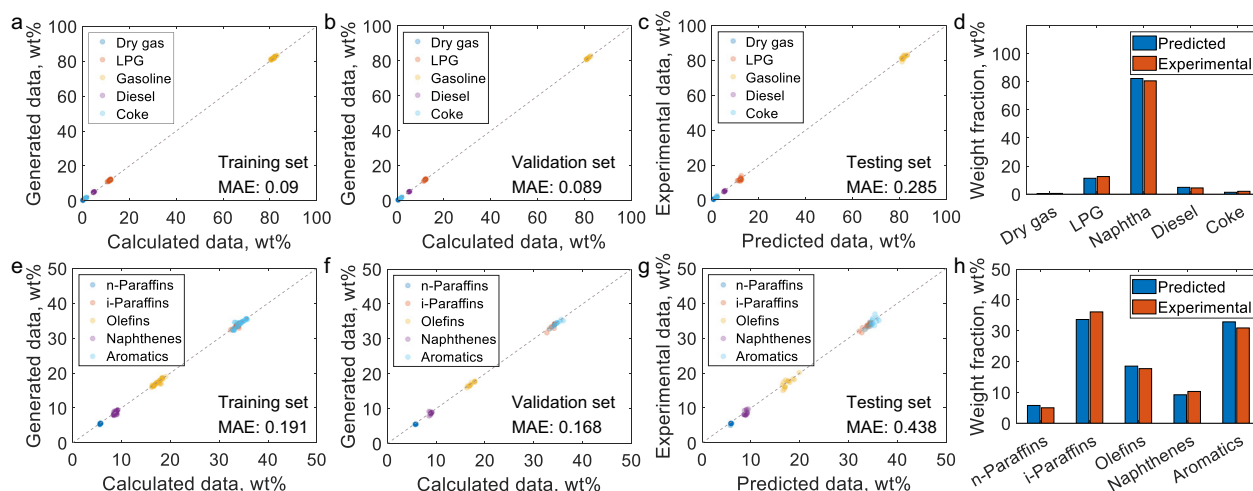


Fig. 8 | Model fine-tuning results for the optimal dataset size. a Parity plot for product yield of naphtha 2 FCC in the training set. **b** Parity plot for product yield of naphtha 2 FCC in the validation set. **c** Parity plot for product yield of naphtha 2 FCC in the testing set. **d** Predicted product yield for naphtha 3 FCC under the following process conditions: Feedstock temperature = 200 °C, Catalyst temperature = 623 °C, Catalyst-to-oil ratio (CTO) = 5.5, and reaction time = 3.6 s. **e** Parity plot for gasoline composition of naphtha 2 FCC in the training set. **f** Parity plot for gasoline composition of naphtha 2 FCC in the validation set. **g** Parity plot for gasoline composition of naphtha 2 FCC in the testing set. **h** Predicted gasoline composition for naphtha 3 FCC under the same process conditions as **d**. The mean absolute error (MAE) for training, validation, and testing sets is reported in the figure. Source data are provided as a Source data file.

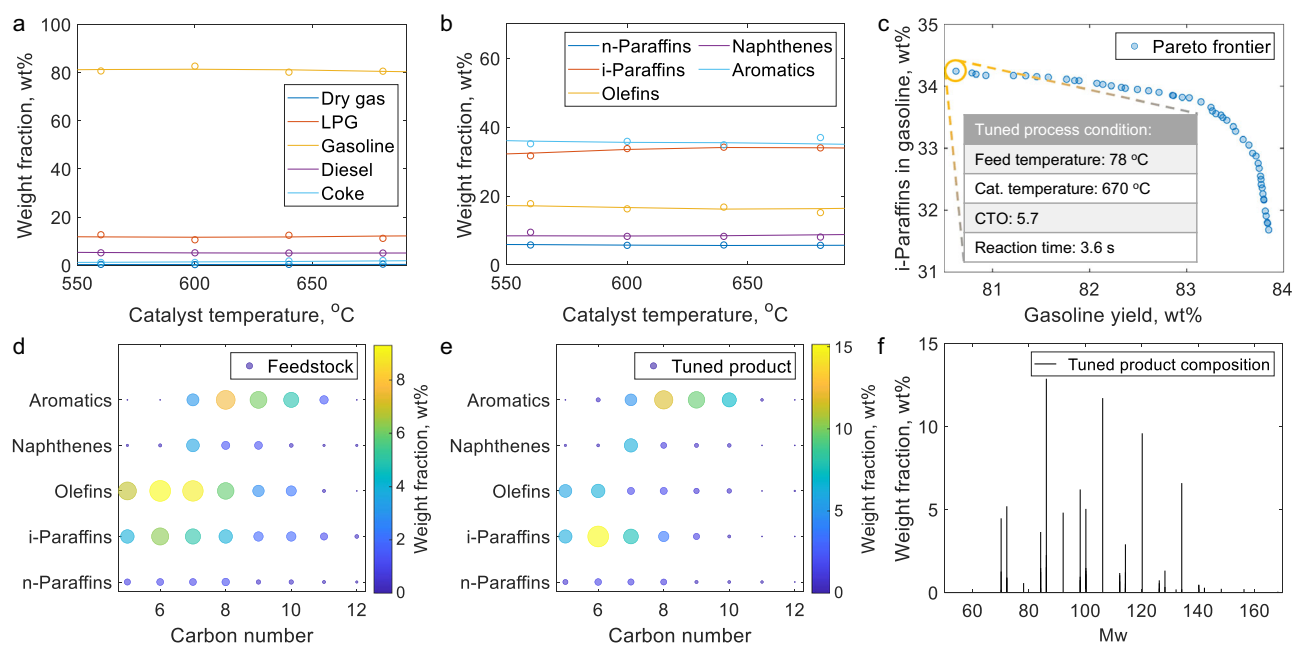


Fig. 9 | Sensitivity analysis and optimal process condition results for naphtha 2 FCC. a Effect of catalyst temperature on product yield. The line represents the predicted values, and the dots denote the experimental data. **b** Effect of catalyst temperature on gasoline composition. The line represents the predicted values, and the dots denote the experimental data. **c** Pareto frontier obtained using the NSGA-II algorithm, with maximum gasoline and i-paraffin yields as optimization objectives. The singled-out point represents the process conditions with the lowest gasoline yield (80.62 wt%) and the highest i-paraffin yield (34.24 wt%) within the Pareto solution set. The corresponding process conditions are listed in the inset

table. **d** Carbon number distribution for naphtha 2. The size and color of the dots indicate the content of each substance in the feedstock. **e** Product carbon number distribution for tuned process condition. The size and color of the dots indicate the content of each substance in the product. Process condition: Feedstock temperature = 78 °C, Catalyst temperature = 670 °C, Catalyst-to-oil ratio (CTO) = 5.7, and Reaction time = 3.6 s. **f** Product molecular composition for tuned process condition. Process condition: Feedstock temperature = 78 °C, Catalyst temperature = 670 °C, CTO = 5.7, and Reaction time = 3.6 s. Source data are provided as a Source data file.

Pilot-scale experiment for naphtha FCC

Pilot-scale experiments of naphtha FCC were performed in a riser reactor, using the same catalyst as laboratory-scale experiments. The main reactor parameters and experimental scheme are illustrated in Supplementary Table 9 and Supplementary Fig. 8, respectively.

Initially, the feedstock and regenerator were preheated to set the temperature according to process conditions. 2 kg/h naphtha was pumped into the bottom of the riser, where it contacted high temperature catalyst from the regenerator. The catalyst circulation rate was calculated based on the CTO. The FCC reaction occurred in the

Table 1 | Tuned process conditions and product distribution

	Experiment data	Tuned data	Model validation
Process conditions			
Feedstock temperature, °C	70	78	78
Catalyst temperature, °C	678	670	670
CTO	5.6	5.7	5.7
Reaction time, s	3.6	3.6	3.6
Bulk properties			
LPG yield, wt%	11.8	12.1	12.2
Gasoline yield, wt%	80.3	80.6	80.5
Coke yield, wt%	2.5	1.8	1.8
i-Paraffin yield, wt%	33.9	34.2	34.2
Olefin yield, wt%	18.2	17.2	17.2
Aromatics yield, wt%	33.9	33.9	33.9
Density at 293 K, °C		0.7428	0.7432
ASTM D86 IBP, °C		26.2	25.8
ASTM D86 50%, °C		84.0	84.8
ASTM D86 FBP, °C		171.7	170.3
Research octane number (RON)		89.8	91.2
Critical pressure (P _c), kPa		3,260.2	3,287.3
Critical temperature (T _c), °C		271.35	269.0

riser. The generated gaseous substances and deactivated catalyst were delivered into the disengager for gas-solid separation. The gaseous phase was collected at the disengager top, while the deactivated catalyst was recycled to the regenerator for coke combustion and regeneration, enabling continuous operation of the unit. The products from the disengager were condensed into gaseous and liquid phases. They were collected and separated in the same method as the laboratory-scale experiments. For the pilot-scale experiments, we selected naphtha 2 and 3 as the feedstock, and the molecular composition of the two naphtha is listed in Supplementary Table 1. Naphtha 2 carried out 15 sets of FCC experiments, while naphtha 3 performed only 1 set of experiments to validate the model accuracy. The process conditions for naphtha 2 and 3 are listed in Supplementary Table 12.

Feedstock and product analysis

The gaseous products from naphtha FCC were quantitatively analyzed using an Agilent 6890 GC. The GC with flame ionization detector (FID) was employed for analyzing hydrocarbons, while the GC with thermal conductivity detector (TCD) was used for analyzing hydrogen. The analytical method used the UOP 539-2012 standard method.

The molecular composition of the naphtha and gasoline fractions was analyzed using an Agilent 6890 GC equipped with an FID, following the ASTM D5134-21 standard method.

The solid product mainly consists of coke deposited on the catalyst, which was analyzed using an HIR-944B high-frequency infrared carbon-sulfur analyzer. The catalyst was subjected to high-temperature combustion, and the CO₂ content in the mixed gas was measured via infrared analysis to determine the coke content.

Data generation from mechanistic model

This study intends to generate a series of laboratory-scale naphtha FCC data to train the ResMLP. These data were generated by a mechanistic model. However, before generating the product data using the

mechanistic model, datasets covering different process conditions and feedstock compositions were prepared. For the process conditions (reaction temperature, WHSV, and CTO), 10,000 data points were randomly sampled according to a normal distribution. These sampled data were able to comprehensively cover the range of process conditions in naphtha FCC processes. For feedstock molecular composition, 10,000 sets of naphtha feedstock data with different molecular compositions were generated by introducing random noise to baseline molecular composition (experimental composition of naphtha 1). Subsequently, these generated data were delivered into the molecular-level kinetic model for naphtha FCC to calculate the corresponding product distributions. Finally, a complete dataset of laboratory-scale naphtha FCC was obtained.

Laboratory-scale hybrid model structure

The proposed hybrid model consists of three multilayer fully connected neural networks, as shown in Fig. 2a. The first network takes process conditions as inputs, including feedstock temperature, reaction temperature, reaction time (WHSV), and CTO. The second neural network takes molecular composition as input, and there are 129 molecules in the naphtha. The outputs of these two networks are then concatenated and delivered to the third network, which predicts the distribution of 129 product molecules. Each hidden layer uses the ReLU activation function, followed by a dropout layer to prevent overfitting.

Naphtha FCC involves a complex molecular reaction network, requiring a deep neural network to capture the nonlinear relationships between feedstock and product molecules. However, increasing network depth often leads to gradient vanishing or explosion. To overcome this limitation, residual connections were introduced between the two fully connected layers to form a ResMLP. Compared to other neural networks incorporating mechanistic models, the ResMLPs allow for efficient forward and backward propagation, and the training time can be significantly reduced. The ResMLP was built using the PyTorch framework in Python. It is easily reproducible for researchers in chemistry and chemical engineering.

Training for laboratory-scale hybrid model

The laboratory-scale hybrid model was trained on 10,000 sets of data generated by the mechanistic model. The data were randomly divided into training and testing sets in the ratio of 8:2. All input data (feedstock molecular composition and process condition) were normalized to the range of −1 to 1. The normalized feedstock molecular composition and laboratory-scale process conditions were fed into Molecule-based ResMLP and Process-based ResMLP, respectively. The outputs of two ResMLPs were concatenated and delivered into the Integrated ResMLP to calculate the product molecular composition. The molecular composition calculated by ResMLP and the mechanistic model were compared in the MSE loss function to compute the error. Then, backpropagation was performed, and the Adam optimizer with weight decay was used to adjust the ResMLP parameters. A learning rate decay strategy was used, with a decay rate of 0.9 and a decay step of 10 epochs. This work applied three error metrics to evaluate the model performance, including R^2 score, MAE, and MSE, as shown in Eqs. (1)–(3). The training was stopped after 300 epochs, and the laboratory-scale hybrid model was obtained.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{\text{cal}} - y_{\text{true}})^2}{\sum_{i=1}^n (y_{\text{cal}} - \bar{y}_{\text{cal}})^2} \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_{\text{cal}} - y_{\text{true}}| \quad (2)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_{\text{cal}} - y_{\text{true}})^2 \quad (3)$$

Where y_{cal} and y_{true} are calculated values by the hybrid model and generated data by the mechanistic model, respectively, and n is the sample number used to train the hybrid model.

Hyperparameter optimization

The proposed network architecture for the hybrid model contains a large number of hyperparameters, such as residual block number, neuron number of each residual block, dropout rate, learning rate, weight decay, and batch size. To efficiently determine optimal hyperparameters, the BO algorithm was employed based on the Optuna package. The model training and hyperparameter optimization were performed based on NVIDIA's 4090 GPU. The hyperparameter optimization process for the laboratory-scale hybrid model is as follows:

Before hyperparameter optimization, the full dataset was randomly split into training (80%) and testing (20%) sets. All input data was normalized to a range of -1 to 1 , and the search space for each hyperparameter was predefined, as shown in Supplementary Table 13. After that, the BO algorithm was called to tune the hyperparameter. Initially, a series of candidate hyperparameters was generated based on the BO algorithm. The training set, testing set, and hyperparameters were input to the ResMLP model for training. When the model training was completed, the R^2 score of the testing set was calculated and selected as the objective function for the BO algorithm. According to the R^2 score, the BO algorithm was repeatedly called to tune the hyperparameters. The optimization process was stopped after 100 iterations. The optimal hyperparameters were obtained from the above 100 iterations when the R^2 score was maximized. In addition to R^2 , other evaluation metrics, such as MAE and root mean square error (RMSE), can also be used as the objective function in the BO algorithm.

Model structure for transfer learning

Due to the similarity of the reaction mechanisms between the laboratory and pilot scales, transfer learning was primarily employed to capture the differences in fluid flow regimes within the reactor and calculate the pilot-scale product distribution. The source domain was laboratory-scale molecular composition data, containing 10,000 samples with 129 features. The target domain was 15 sets of pilot-scale bulk property data for naphtha 2 FCC. Due to data discrepancies between the source domain and target domain, directly calculating the pilot-scale product distribution using conventional transfer learning posed a challenge. To address this, this work incorporated some bulk property equations into the loss functions, including fraction yield and gasoline composition. The fraction yields were divided primarily based on the molecular boiling points as follows:

Dry gas: $0 \text{ K} < \text{Molecular boiling point} < 220 \text{ K}$

LPG: $220 \text{ K} < \text{Molecular boiling point} < 280 \text{ K}$

Gasoline: $280 \text{ K} < \text{Molecular boiling point} < 473 \text{ K}$

Diesel: $473 \text{ K} < \text{Molecular boiling point} < 623 \text{ K}$

Coke: $623 \text{ K} < \text{Molecular boiling point}$, and molecules with more than 5 aromatic rings and less than 2 carbons in the side chain.

The molecular boiling points were calculated by the group contribution method, followed by dividing them according to the scheme described above. Once the fraction yields were calculated, gasoline was further classified into n-paraffins, i-paraffins, olefins, naphthenes, and aromatics based on molecular structure. The specific classification scheme is as follows:

n-Paraffins: Molecules without branched chains, double bonds, and rings

i-Paraffins: Molecules with branched chains and without double bonds and rings

Olefins: Molecules with double bonds and without rings

Naphthenes: Molecules with naphthenic rings and without aromatic rings

Aromatics: Molecules with aromatic rings

According to the above method, the gasoline composition can be calculated. When the property information is incorporated into the hybrid model, the calculated property data can be directly compared to experimental data during model training.

Parameter fine-tuning in transfer learning

Although 15 sets of pilot-scale experimental data were obtained, training a robust model directly with such a limited dataset was a challenging task. To overcome this limitation, data augmentation was performed, and pilot-scale product data were generated through interpolation within the range of available pilot data. The augmented data were randomly split into training and validation sets with an 8:2 ratio, while the original 15 sets of experimental data were reserved as the testing set and excluded from training. All input data (feedstock molecular composition and process condition) were scaled to the range of -1 to 1 . The normalized feedstock molecular composition and pilot-scale process conditions were input into Molecule-based ResMLP and Process-based ResMLP, respectively. Since the feedstock molecular composition and catalyst remained almost unchanged during the scale-up from laboratory to pilot plant, the network parameter of Molecule-based ResMLP was frozen, while the Process-based and Integrated ResMLPs were fine-tuned. Then, the calculated molecular composition was input into the bulk property equations to estimate the product yield and gasoline composition. The calculated and measured bulk properties were compared in the MSE loss function to compute the error. Similar to training the laboratory-scale hybrid model, backpropagation was performed, and the Adam optimizer with weight decay was used to fine-tune the ResMLP parameters. The performance of the transfer learning was evaluated using the R^2 score, MAE, and MSE. The fine-tuning was stopped after 300 epochs, and the pilot-scale hybrid model was obtained.

Moreover, the hyperparameters for transfer learning were also tuned by the BO algorithm. Since the neural network architecture has been determined, only four hyperparameters need to be tuned, including dropout rate, learning rate, weight decay, and batch size. Their ranges are listed in Supplementary Table 14. The hyperparameter optimization process was the same as the laboratory-scale hybrid model.

Process condition optimization by NSGA-II

This study utilized the NSGA-II to optimize the process conditions of the pilot-scale unit, including feedstock temperature, regenerated catalyst temperature, and CTO. The maximum and minimum values of these conditions, derived from pilot experiments, were defined as constraints, as shown in Eqs. (4)–(6).

$$T_{\text{feed}}^{\min} \leq T_{\text{feed}} \leq T_{\text{feed}}^{\max} \quad (4)$$

$$T_{\text{cat}}^{\min} \leq T_{\text{cat}} \leq T_{\text{cat}}^{\max} \quad (5)$$

$$\text{CTO}^{\min} \leq \text{CTO} \leq \text{CTO}^{\max} \quad (6)$$

Where T_{cat} and T_{feed} are regenerated catalyst temperature and feedstock temperature, respectively. T_{cat}^{\min} and T_{cat}^{\max} are the minimum and maximum temperature for the regenerated catalyst. T_{feed}^{\min} and T_{feed}^{\max} is the minimum and maximum temperature for feedstock, and CTO^{\min} and CTO^{\max} are Minimum and Maximum CTO, respectively.

After that, the NSGA-II algorithm was coupled with the hybrid model to optimize the process condition. NSGA-II initially generated a range of process conditions within the defined upper and lower bounds. These conditions were input into the trained pilot-scale hybrid model to calculate the product yield and gasoline composition. The objective functions were to maximize gasoline and i-paraffin yields, as defined in Eqs. (7) and (8). Then, calculated yields of gasoline and i-paraffins were fed back into the NSGA-II, and the algorithm iteratively searched for tuned process conditions. After 50 iterations, the Pareto-optimal solution was obtained, and the optimal process conditions were identified from the Pareto front. The calculation process is shown in Supplementary Fig. 9.

$$\min y_{\text{gasoline}} = - \sum_{i=1}^n x_i (280 \text{ K} < \text{Molecular boiling point} < 473 \text{ K}) \quad (7)$$

$$\min y_{\text{i-paraffin}} = - \sum_{i=1}^n x_i (\text{i-paraffin molecules}) \quad (8)$$

Where x_i is the molecule mass fraction for species i . y_{gasoline} and $y_{\text{i-paraffin}}$ are the gasoline fraction yield and i-Paraffin yield in the gasoline fraction, respectively. n is the number of molecules.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used for model training and prediction in this study are publicly available at: <https://github.com/chenzhengyuAI/complex-reaction-system-scale-up>⁴⁸. Moreover, Source data are provided with this paper.

Code availability

The code of the hybrid model and all scripts used to train and predict product distribution are publicly available at: <https://github.com/chenzhengyuAI/complex-reaction-system-scale-up>⁴⁸.

References

- Marin, G. B., Galvita, V. V. & Yablonsky, G. S. Kinetics of chemical processes: from molecular to industrial scale. *J. Catal.* **404**, 745–759 (2021).
- Froment, G. F. Kinetic modeling of acid-catalyzed oil refining processes. *Catal. Today* **52**, 153–163 (1999).
- Wei, W. et al. Computer aided kinetic modeling with KMT and KME. *Fuel Process. Technol.* **89**, 350–363 (2008).
- Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).
- Quann, R. J. & Jaffe, S. B. Structure-oriented lumping: describing the chemistry of complex hydrocarbon mixtures. *Ind. Eng. Chem. Res.* **31**, 2483–2497 (1992).
- Quann, R. J. & Jaffe, S. B. Building useful models of complex reaction systems in petroleum refining. *Chem. Eng. Sci.* **51**, 1615–1635 (1996).
- Peng B. *Molecular modelling of petroleum processes* (The University of Manchester, 1999).
- Feng, S. et al. Molecular composition modelling of petroleum fractions based on a hybrid structural unit and bond-electron matrix (SU-BEM) framework. *Chem. Eng. Sci.* **201**, 145–156 (2019).
- Feng, W., Vynckier, E. & Froment, G. F. Single event kinetics of catalytic cracking. *Ind. Eng. Chem. Res.* **32**, 2997–3005 (1993).
- Quintana-Solórzano, R., Thybaut, J. W., Galtier, P. & Marin, G. B. Simulation of an industrial riser for catalytic cracking in the presence of coking using Single-Event MicroKinetics. *Catal. Today* **150**, 319–331 (2010).
- Standl, S. et al. Single-event kinetic model for methanol-to-olefins (MTO) over ZSM-5: fundamental kinetics for the olefin co-feed reactivity. *Chem. Eng. J.* **402**, 126023 (2020).
- Horton, S. R., Mohr, R. J., Zhang, Y., Petrocelli, F. P. & Klein, M. T. Molecular-level kinetic modeling of biomass gasification. *Energy Fuels* **30**, 1647–1661 (2016).
- Horton, S. R. et al. Molecular-level kinetic modeling of resid pyrolysis. *Ind. Eng. Chem. Res.* **54**, 4226–4235 (2015).
- Zhou, X., Li, W., Mabon, R. & Broadbelt, L. J. A mechanistic model of fast pyrolysis of hemicellulose. *Energy Environ. Sci.* **11**, 1240–1260 (2018).
- Ghosh, P., Andrews, A. T., Quann, R. J. & Halbert, T. R. Detailed Kinetic model for the hydro-desulfurization of FCC naphtha. *Energy Fuels* **23**, 5743–5759 (2009).
- Christensen, G., Apelian, M. R., Hickey, K. J. & Jaffe, S. B. Future directions in modeling the FCC process: an emphasis on product quality. *Chem. Eng. Sci.* **54**, 2753–2764 (1999).
- Liu, L. *Molecular Characterisation and Modelling for Refining Processes* (The University of Manchester, 2015).
- Chen, Z. et al. Molecular-level kinetic modeling of heavy oil fluid catalytic cracking process based on hybrid structural unit and bond-electron matrix. *AIChE J.* **67**, e17027 (2021).
- Neurock, M., Libanati, C., Nigam, A. & Klein, M. T. Monte carlo simulation of complex reaction systems: molecular structure and reactivity in modelling heavy oils. *Chem. Eng. Sci.* **45**, 2083–2088 (1990).
- Vernuccio, S. & Broadbelt, L. J. Discerning complex reaction networks using automated generators. *AIChE J.* **65**, e16663 (2019).
- Van Geem, K. M. et al. Automatic reaction network generation using RMG for steam cracking of n-hexane. *AIChE J.* **52**, 718–730 (2006).
- Klein M. T., Gang H., Bertolacini R., Broadbelt L. J. & Kumar, A. *Molecular Modeling in Heavy Hydrocarbon Conversions* (CRC Press, 2005).
- Zhu, F. X. X. & Xu, L. Integrating multiscale modeling and optimization for sustainable process development. *Chem. Eng. Sci.* **254**, 117619 (2022).
- Dewachtere, N. V., Santaella, F. & Froment, G. F. Application of a single-event kinetic model in the simulation of an industrial riser reactor for the catalytic cracking of vacuum gas oil. *Chem. Eng. Sci.* **54**, 3653–3660 (1999).
- Lu, L. et al. EMMS-based discrete particle method (EMMS-DPM) for simulation of gas–solid flows. *Chem. Eng. Sci.* **120**, 67–87 (2014).
- Gao, J., Xu, C., Lin, S., Yang, G. & Guo, Y. Advanced model for turbulent gas–solid flow and reaction in FCC riser reactors. *AIChE J.* **45**, 1095–1113 (1999).
- Kochkov, D. et al. Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. USA* **118**, e2101784118 (2021).
- Fisher, O. J. et al. Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Comput. Chem. Eng.* **140**, 106881 (2020).
- Sharma, N. & Liu, Y. A. A hybrid science-guided machine learning approach for modeling chemical processes: a review. *AIChE J.* **68**, e17609 (2022).
- Schweidtmann, A. M. Generative artificial intelligence in chemical engineering. *Nat. Chem. Eng.* **1**, 193–193 (2024).
- Dobbelaere, M. R., Plehiers, P. P., Van de Vijver, R., Stevens, C. V. & Van Geem, K. M. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. *Engineering* **7**, 1201–1211 (2021).
- Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).

33. Chen R. T. Q., Rubanova Y., Bettencourt J. & Duvenaud, D. Neural ordinary differential equations. In *Proc. 32nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2018).
34. Yin, J., Li, J., Karimi, I. A. & Wang, X. Generalized reactor neural ODE for dynamic reaction process modeling with physical interpretability. *Chem. Eng. J.* **452**, 139487 (2023).
35. Hough, B. R., Beck, D. A., Schwartz, D. T. & Pfaendtner, J. Application of machine learning to pyrolysis reaction networks: reducing model solution time to enable process optimization. *Comput. Chem. Eng.* **104**, 56–63 (2017).
36. Hua, F., Fang, Z. & Qiu, T. Application of convolutional neural networks to large-scale naphtha pyrolysis kinetic modeling. *Chin. J. Chem. Eng.* **26**, 2562–2572 (2018).
37. Chen, Z. et al. Molecular-Level modeling for naphtha olefin reduction in FCC subsidiary riser: from laboratory reactor to pilot plant. *Chem. Eng. J.* **437**, 135429 (2022).
38. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2021).
39. Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D. & Lió, P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat. Commun.* **15**, 1517 (2024).
40. Wu, H. & Zhao, J. Fault detection and diagnosis based on transfer learning for multimode chemical processes. *Comput. Chem. Eng.* **135**, 106731 (2020).
41. Xiao, M., Hu, C. & Wu, Z. Modeling and predictive control of non-linear processes using transfer learning method. *AIChE J.* **69**, e18076 (2023).
42. Cui, C. et al. Computer-aided gasoline compositional model development based on GC-FID analysis. *Energy Fuels* **32**, 8366–8373 (2018).
43. Korre, S. C., Neurock, M., Klein, M. T. & Quann, R. J. Hydrogenation of polynuclear aromatic hydrocarbons. 2. quantitative structure/reactivity correlations. *Chem. Eng. Sci.* **49**, 4191–4210 (1994).
44. Korre, S. C., Klein, M. T. & Quann, R. J. Polynuclear aromatic hydrocarbons hydrogenation. 1. Experimental reaction pathways and kinetics. *Ind. Eng. Chem. Res.* **34**, 101–117 (1995).
45. Mochida, I. Yoneda Y. Linear free energy relationships in heterogeneous catalysis: II. Dealkylation and isomerization reactions on various solid acid catalysts. *J. Catal.* **7**, 393–396 (1967).
46. Xu, Y. & Cui, S. A novel fluid catalytic cracking process for maximizing iso-paraffins: from fundamentals to commercialization. *Front. Chem. Sci. Eng.* **12**, 9–23 (2018).
47. Zhu, X. et al. Conceptual fluid catalytic cracking process with the additional regenerated catalyst circulation path for gasoline reprocessing and upgrading with minimum loss. *Energy Fuels* **34**, 235–244 (2020).
48. Zhengyu, C., Yongqing, X., Chunming, X. & Linzhou Z. Scale-up of complex molecular reaction system by hybrid mechanistic modeling and deep transfer learning. (2025). Repository name: complex-reaction-system-scale-up. <https://doi.org/10.5281/zenodo.16810762>

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 22222815 to L.Z., 22441040 to L.Z. and 22021004 to C.X.), China Postdoctoral Science Foundation (No. 2024M753610 to Z.C.), and Science Foundation of China University of Petroleum, Beijing (Nos. 2462024XKBH004 to Z.C., 2462025BJRC007 to Z.C., and 2462023QNXZ004 to L.Z.).

Author contributions

Z.C. and L.Z. contributed to the conceptualization and methodology of the hybrid model; Z.C. and Y.X. implemented the algorithm under the supervision of L.Z. and C.X.; Z.C. analyzed data and wrote the paper; All authors reviewed and discussed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-63982-2>.

Correspondence and requests for materials should be addressed to Linzhou Zhang.

Peer review information *Nature Communications* thanks Magda Helena Barecka, Cameron J. Brown, and the other anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025