
Predicting the Performance of Fluid Catalytic Cracking (FCC) or Hydrocracking Catalysts Using Deep Learning Models

Mahsa Nadimpiur¹

1. Department of Chemistry, Faculty of Basic Sciences, Shahid Chamran University, Ahvaz, Iran

Abstract .

Fluid Catalytic Cracking (FCC) units are critical yet operationally complex systems characterized by severe non-linearities, strong phase interactions, and harsh thermodynamic environments, making real-time monitoring and product yield prediction challenging. While traditional first-principles kinetic and hydrodynamic models provide mechanistic fidelity, their high computational cost limits real-time closed-loop control. Conversely, purely data-driven soft sensors often suffer from poor generalization and lack interpretability. This study presents a comprehensive state-of-the-art review and comparative framework integrating multi-lump kinetic models, multiphase hydrodynamic intensification, and advanced deep learning architectures—including spatial-temporal neural networks, physics-guided neural networks (PGNN), and explainable artificial intelligence (XAI). By synthesizing mechanistic multi-phase interactions in the reactor-regenerator loop with data-driven soft sensing methodologies, we outline critical design trade-offs in prediction accuracy, physical consistency, and computational overhead. Furthermore, feature attribution frameworks utilizing Shapley Additive Explanations (SHAP) are evaluated to ensure transparent, trustworthy decision-making for plant-wide optimization. This work establishes a robust blueprint for developing next-generation, physics-informed digital twins and intelligent soft sensors in sustainable petroleum refining.

Keywords: Fluid Catalytic Cracking (FCC); Soft Sensor; Physics-Guided Neural Networks; Kinetic Modeling; Explainable AI (XAI).

Introduction

Fluid catalytic cracking is a central conversion technology in petroleum refining because it transforms relatively heavy hydrocarbon streams into higher-value products, particularly gasoline-range hydrocarbons and light olefins. The process simultaneously involves catalytic reactions, rapid heat transfer, gas–solid contacting, catalyst circulation, coke formation, and regeneration. Consequently, FCC performance cannot be represented adequately by a single isolated reaction or unit operation; it is more appropriately viewed as a tightly coupled reactor–regenerator system in which reaction and transport phenomena interact over multiple spatial and temporal scales (Hsu & Robinson, 2019a, 2019b; Sadeghbeigi, 2020).

The practical importance of FCC optimization is also associated with the changing objectives of contemporary refineries. Product specifications, feedstock variability, energy consumption, emissions constraints, and demands for higher-value olefins can shift the operating window of the unit. Kinetic modeling remains valuable because it provides a mechanistic representation of conversion and selectivity, while hydrodynamic modeling is required to represent gas–solid flow, catalyst distribution, mixing, and heat and mass transfer. Studies based on six-lump and more detailed lumping strategies demonstrate that the number and definition of kinetic lumps can materially affect the ability of a model to describe heavy-oil conversion and product distributions (Chen et al., 2020; Olafadehan et al., 2019).

Despite their physical value, mechanistic models can be computationally demanding when coupled with detailed reactor and regenerator descriptions. Equivalent-reactor-network formulations have therefore been investigated as a means of representing the integrated FCC loop while retaining a tractable computational structure (Du et al., 2019). At the same time, continuum and particle-based descriptions of gas–solid flow have been developed to capture the hydrodynamics that influence catalyst residence time, solids distribution, mixing, and regeneration performance (Wang, 2020; Yang et al., 2021a, 2021b). These developments illustrate a fundamental modeling challenge: increasing physical fidelity can improve mechanistic realism but may reduce suitability for real-time optimization.

Data-driven soft sensors offer a complementary strategy. Rather than solving the complete process model online, a soft sensor learns a mapping between readily available process measurements and difficult-to-measure quality or performance variables. The process-industry literature has established soft sensing as an important approach for inferential monitoring, while more recent deep-learning methods seek to exploit nonlinear, temporal, and multivariable structure in industrial data (Kadlec et al., 2009; Sun & Ge, 2021). Applications in refinery operations further show the practical relevance of machine-learning soft sensors for predicting process variables that may otherwise require laboratory analysis or specialized instrumentation (Ferreira et al., 2022).

However, a purely data-driven model can learn correlations that are statistically useful within a training domain but physically implausible outside it. This concern becomes particularly important for FCC applications because feed properties, catalyst activity, operating conditions, and reactor–regenerator interactions can change over time. Data-driven process models also face challenges related to data quality, sensor faults, process drift, variable correlation, and limited coverage of abnormal operating regimes (Fisher et al., 2020). Physics-guided learning provides a pathway for incorporating process knowledge into flexible neural architectures. Recent work on

physics-guided graph learning, for example, demonstrates how structural process information can be integrated into soft-sensor design (Liu et al., 2024).

Explainability is an additional requirement for industrial deployment. A highly accurate prediction that cannot be interpreted by process engineers may be difficult to trust, validate, or use in operational decision-making. Explainable artificial intelligence approaches based on feature attribution can expose the contribution of input variables to individual predictions. Shapley-value-based methods such as SHAP have consequently been investigated for interpreting machine-learning models in engineering and process applications (Bhakte et al., 2022; Sharma Timilsina et al., 2024).

Against this background, the present article develops a comparative framework for predicting FCC or hydrocracking catalyst performance using deep learning. The emphasis is not on replacing mechanistic knowledge with machine learning, but on identifying how kinetic, hydrodynamic, temporal, graph-based, physics-guided, and explainability components can be combined into a coherent predictive architecture. The article reviews the relevant process foundations, kinetic representations, hydrodynamic considerations, soft-sensor methodologies, deep-learning architectures, physics-guided strategies, and XAI methods, and then proposes a research framework suitable for future experimental or industrial validation.

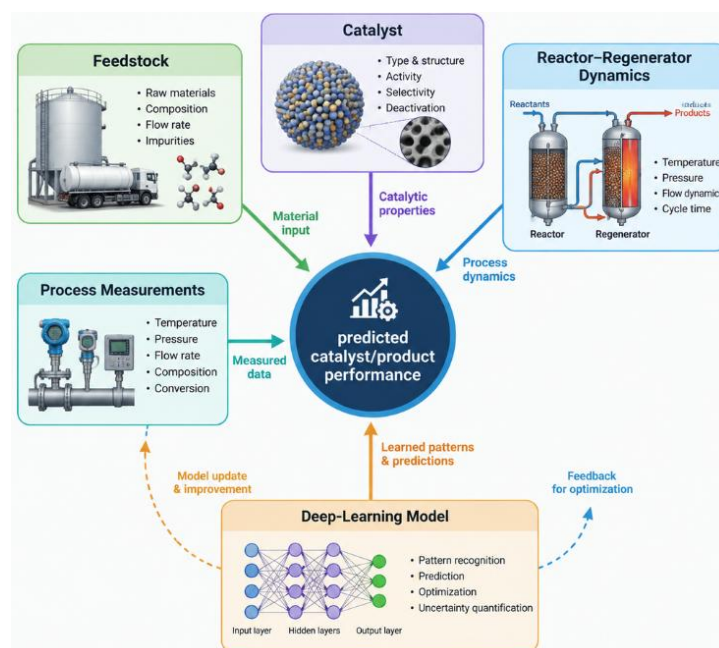


Figure 1. Conceptual relationship among feedstock, catalyst, reactor–regenerator dynamics, process measurements, deep-learning model, and predicted catalyst/product performance.

2. FCC and Hydrocracking as Complex Catalytic Conversion Systems

2.1. FCC process fundamentals

FCC is based on the catalytic conversion of heavy hydrocarbon feedstocks over acidic catalysts, with the riser reactor providing the primary reaction environment and the regenerator restoring catalyst activity by removing deposited coke. The rapid circulation of catalyst between the reactor and regenerator makes heat balance and catalyst activity inseparable from reaction performance. General descriptions of cracking chemistry emphasize the importance of feed composition, catalyst properties, reaction temperature, contact conditions, and product distribution in determining conversion and selectivity (Hsu & Robinson, 2019a; Speight, 2019).

From an operational perspective, FCC performance is affected by reactor temperature, catalyst-to-oil ratio, residence time, feed quality, catalyst activity, coke yield, and regeneration conditions. The practical optimization of these variables requires simultaneous consideration of conversion, gasoline selectivity, light-olefin production, dry-gas formation, coke deposition, and energy balance (Sadeghbeigi, 2020).

2.2. Hydrocracking context

Hydrocracking differs from FCC in its reliance on hydrogen and bifunctional catalytic chemistry, combining acidic cracking functions with hydrogenation–dehydrogenation functionality. Although the process environment is different, several methodological challenges relevant to predictive modeling are shared: strong nonlinearities, interacting process variables, catalyst-state dependence, feedstock variability, and the need to infer difficult-to-measure performance variables. Accordingly, the machine-learning framework discussed here can be adapted to hydrocracking when suitable catalyst, feed, hydrogen, and reactor variables are available. The FCC literature is used as the principal process case because the supplied reference set is substantially concentrated on FCC kinetics, hydrodynamics, regeneration, and soft sensing.

3. Catalyst Performance and Kinetic Representation

3.1. Lumped kinetic models

Detailed hydrocarbon reaction networks can contain an impractically large number of individual species and reactions for routine process simulation. Lumped kinetic models reduce this complexity by grouping chemically similar components into pseudo-components or lumps. Six-lump models have been used to simulate vacuum-gas-oil cracking in riser reactors, providing a tractable representation of conversion and product formation (Olafadehan et al., 2019). More detailed schemes, including a 12-lump kinetic model, can distinguish a wider range of feed and product categories and can be designed around objectives such as gasoline cleaning and enhanced light-olefin production (Chen et al., 2020).

The choice of lumping resolution is therefore a modeling trade-off. A coarse model is computationally attractive and can be easier to calibrate, but it may fail to resolve product pathways that become important under changing feedstocks or operating targets. A more detailed model can provide richer mechanistic information but introduces additional parameters and potential identifiability problems. For a hybrid deep-learning framework, the kinetic model need not necessarily be executed in full during every prediction; instead, its outputs, constraints, or latent states can serve as physics-guided information for the learning system.

Table 1. Comparison of kinetic-modeling levels for FCC prediction: lump definition, physical fidelity, data requirements, computational burden, and suitability for hybrid learning.

Kinetic modeling level	Lump definition / representation	Physical fidelity	Data & parameter requirements	Computational burden	Suitability for hybrid deep learning
Six-lump kinetic model	Feed and products are represented using a limited number of pseudo-components, typically distinguishing major feed and product families such as heavy oil, gasoline, light gases, coke, and related fractions.	Moderate; captures the principal conversion pathways but provides limited resolution of individual product families.	Relatively low to moderate; requires kinetic parameters for a limited reaction network and process operating data for calibration.	Low–moderate; generally suitable for rapid simulation and generation of mechanistic features.	High; useful as a physics-based baseline, auxiliary predictor, or source of kinetic constraints for neural networks.
Multi-lump kinetic model (e.g., 10–12 lumps)	Hydrocarbon feed and products are divided into a larger number of pseudo-components to distinguish important reaction pathways and product groups.	Moderate–high; provides improved representation of gasoline, light olefins, heavy fractions, and other product pathways.	Moderate–high; requires additional kinetic parameters and sufficiently informative experimental or plant data.	Moderate; greater number of states and kinetic relationships increases computational requirements.	Very high; suitable for generating physics-informed features, auxiliary targets, and mechanistic loss terms.
Detailed lumped kinetic model	Uses a relatively large number of pseudo-components and reaction pathways to represent feed conversion and product formation with greater chemical resolution.	High; captures more detailed product distributions and reaction interactions.	High; requires extensive calibration data and reliable estimation of a larger parameter set.	Moderate–high; suitable primarily for offline simulation, model development, and surrogate-data generation.	High, particularly for physics-guided neural networks (PGNNs) and residual-learning frameworks.
Integrated reactor–regenerator kinetic model	Couples riser reaction kinetics with catalyst regeneration, coke combustion, catalyst	High at the process-system level because reactor and regenerator interactions	High; requires kinetic, thermal, catalyst-circulation, and operating data from both reactor and regenerator sections.	High compared with isolated riser models because of the coupled reactor–regenerator equations.	Very high; provides mechanistic latent states and physical constraints for plant-wide deep-

	circulation, and associated heat/material balances.	are explicitly represented.	learning models.
Kinetic–hydrodynamic coupled model	Combines reaction kinetics with gas–solid flow, catalyst distribution, mixing, residence-time, and heat/mass-transfer effects.	Very high; represents both reaction and multiphase transport phenomena.	Very high; requires kinetic parameters, hydrodynamic parameters, geometry, operating data, and often simulation/experimental information.

3.2. Catalyst activity and deactivation

Catalyst performance cannot be treated as a static property because catalyst circulation and regeneration continuously alter the state of the catalytic material. Coke deposition reduces available catalytic functionality, while regeneration restores activity through coke combustion. Catalyst formulation and structure can also influence product selectivity. For example, catalyst-promoter studies have investigated modifications intended to alter gasoline composition and selectivity (Liu et al., 2020), illustrating that catalyst descriptors can be relevant predictive features rather than merely background information.

A predictive model should therefore distinguish between instantaneous operating variables and slowly varying catalyst-state variables. Catalyst age, estimated activity, regenerated catalyst temperature, coke loading, circulation rate, and catalyst-to-oil ratio can be represented as dynamic inputs or latent states. This distinction is important for avoiding a common soft-sensor limitation in which a model fits short-term correlations but does not adequately represent long-term process drift.

4. Reactor–Regenerator Integration and Multiphase Hydrodynamics

4.1. Integrated reactor–regenerator modeling

The FCC loop is intrinsically integrated because the regenerator supplies hot regenerated catalyst to the reactor while the reactor returns spent catalyst containing coke. Consequently, reactor conversion and regenerator performance are coupled through catalyst circulation, temperature, coke balance, and heat transfer. An integrated reactor–regenerator model based on an equivalent reactor network has been proposed to represent this coupling while reducing the complexity associated with fully detailed descriptions (Du et al., 2019).

For predictive machine learning, the reactor–regenerator structure provides a natural graph representation. Nodes can correspond to the riser, stripper, regenerator, catalyst cooler, cyclones, and other major process sections, while edges represent material, energy, or catalyst flows. Such a representation can complement conventional feature vectors and provide structural information that is difficult to encode using an unordered list of measurements. Knowledge-enhanced graph learning has recently been explored for multiunit industrial plant-wide monitoring, supporting the

broader argument that process topology can be embedded in learning architectures (Wu et al., 2024).

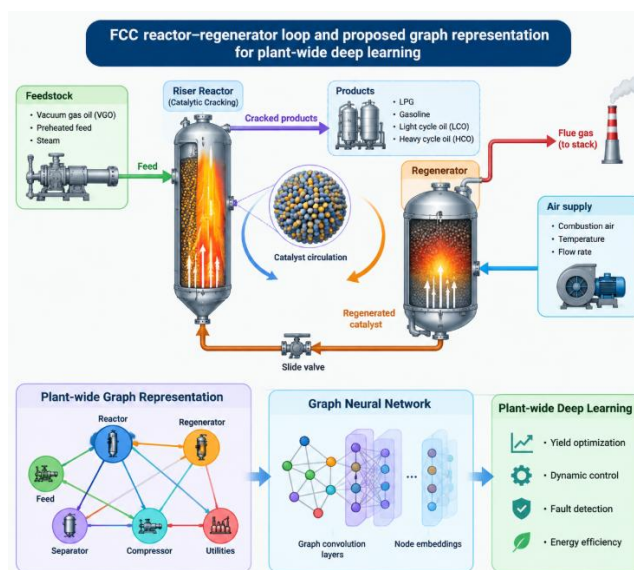


Figure 2. FCC reactor-regenerator loop and proposed graph representation for plant-wide deep learning.

4.2. Gas-solid flow and hydrodynamic effects

Gas-solid hydrodynamics influence catalyst distribution, residence time, mixing, phase contacting, pressure drop, and heat and mass transfer. Continuum theories provide a general framework for describing dense gas-solid systems, while computational approaches such as MP-PIC and CPFD can be used to investigate detailed solids distribution and the influence of internal structures (Wang, 2020; Yang et al., 2019; Yang et al., 2021a, 2021b).

Regenerator performance is particularly sensitive to the distribution of spent catalyst and internal structures. MP-PIC simulations have been used to examine the effects of spent-catalyst distribution and horizontal baffles on industrial FCC regenerator performance (Yang et al., 2021a, 2021b). Related work has considered the forces acting on baffles in gas-solid fluidized beds, emphasizing the mechanical and hydrodynamic complexity of internal components (Liu et al., 2022).

Catalyst cooling and heat removal are also relevant because the catalyst circulation loop must manage the thermal duty generated during regeneration. Experimental and computational studies of external catalyst coolers have investigated particle distribution and heat-transfer behavior under different inlet configurations (Li et al., 2020a, 2020b). These observations suggest that hydrodynamic variables can be valuable auxiliary inputs for learning-based prediction even when a full computational-fluid-dynamics simulation is not performed online.

4.3. Multiphase intensification and regeneration

Regeneration intensification is an important pathway for improving FCC performance because regeneration affects catalyst activity, heat balance, emissions, and catalyst circulation. A state-of-the-art review of regeneration intensification technologies highlights the variety of strategies

available for improving the regeneration step and the associated hydrodynamic and heat-transfer challenges (Oloruntoba et al., 2022). Distributor design and gas distribution can also modify fluidization and the coupling between hydrodynamics and reaction, reinforcing the need to consider multiphase structure in process modeling (Oloruntoba et al., 2022).

5. Soft Sensors for Complex Refining Processes

5.1. Conventional soft-sensor concept

A soft sensor estimates a difficult-to-measure process variable from other available measurements. In process industries, soft sensors are particularly useful when laboratory measurements are delayed, expensive, infrequent, or unavailable at the desired sampling frequency. The foundational literature identifies model-based and data-driven approaches and emphasizes the importance of model maintenance under changing process conditions (Kadlec et al., 2009).

For refinery applications, soft sensors can infer product quality, composition, conversion, or other performance indicators from online process variables. A machine-learning soft sensor developed for an oil-refinery distillation column illustrates how nonlinear data-driven models can be applied to refinery process inference (Ferreira et al., 2022). For FCC, analogous targets could include conversion, gasoline yield, light-olefin yield, coke yield, catalyst activity, or selected product-quality indicators.

5.2. Challenges in industrial data

Industrial process datasets differ substantially from laboratory datasets. Measurements may contain missing values, sensor bias, asynchronous sampling, maintenance periods, operating transitions, and long periods in which the process remains near a limited number of operating regimes. Data-driven model development must therefore consider data representativeness, temporal dependence, leakage between training and test sets, process drift, and the physical meaning of candidate predictors (Fisher et al., 2020).

Deep learning can extract hierarchical representations from high-dimensional process data, but increased model capacity does not automatically guarantee better industrial generalization. A model trained on one feedstock or operating campaign may perform poorly when catalyst properties or operating conditions shift. Therefore, validation should be performed using temporally separated data and, where possible, independent operating campaigns rather than random splitting alone.

6. Deep Learning Architectures for Catalyst-Performance Prediction

6.1. Multilayer neural networks

Feed-forward neural networks provide a baseline nonlinear mapping from process measurements to target variables. Their principal advantage is simplicity and computational efficiency at inference time. However, conventional multilayer architectures do not explicitly exploit temporal ordering or plant topology. For strongly dynamic FCC systems, these limitations motivate architectures that encode time, spatial relationships, or both.

6.2. Convolutional and multichannel architectures

Convolutional neural networks can identify local patterns in ordered process measurements. For dynamic processes, multichannel convolutional structures can capture complementary temporal

patterns across groups of variables and have been investigated for soft-sensor modeling (Yuan et al., 2020). Adaptive multiscale convolutional approaches have also been proposed for chemical-process fault diagnosis, illustrating the usefulness of multiple receptive-field scales for complex process signals (Qin & Zhao, 2022).

6.3. Spatiotemporal learning

Industrial processes contain both temporal dependencies and structural relationships among units and variables. A local spatiotemporal structure-preserving deep-learning approach has been proposed for soft sensing in complex industrial processes, highlighting the benefit of retaining local process structure during representation learning (Wang et al., 2024). For FCC, temporal windows can encode catalyst circulation and process dynamics, while spatial or unit-level structure can encode the relationships between reactor, regenerator, cooler, and separation sections.

6.4. Graph neural networks

Graph-based learning provides a natural mechanism for representing process topology. A plant can be modeled as a graph in which process units are nodes and streams or physical relationships are edges. Knowledge-enhanced distributed graph autoencoders have been investigated for multiunit plant-wide monitoring, demonstrating the potential of graph representations for capturing distributed industrial dependencies (Wu et al., 2024). Physics-guided graph learning has further been proposed specifically for chemical-process soft sensors (Liu et al., 2024).

For an FCC application, node features could include riser temperature, regenerator temperature, catalyst circulation, pressure, oxygen concentration, and other unit-specific measurements, while edge features could represent mass, energy, or catalyst flows. This representation may reduce the tendency of a conventional dense network to treat physically connected and unrelated variables as equivalent.

7. Physics-Guided Neural Networks and Hybrid Modeling

Physics-guided learning aims to combine the flexibility of data-driven models with information supplied by physical laws, conservation equations, kinetic relationships, or known process constraints. The underlying principle is broader than simply adding a mechanistic model as another input: physical knowledge can influence model architecture, loss functions, latent states, training targets, or prediction constraints.

For FCC, candidate physical constraints include mass conservation, energy balance, non-negativity of yields and concentrations, catalyst circulation relationships, coke balance, and qualitative relationships between reaction temperature and conversion. A kinetic model can also generate auxiliary targets such as estimated conversion rates or reaction-pathway contributions. The neural network then learns residual relationships not captured by the mechanistic model.

One practical formulation is a hybrid prediction of the form $\hat{y} = f_{\text{phys}}(x, \theta) + f_{\text{DL}}(x, z)$, where f_{phys} represents a mechanistic baseline and f_{DL} learns a correction term. An alternative is to define a composite loss $L = L_{\text{data}} + \lambda_{\text{phys}} L_{\text{phys}} + \lambda_{\text{reg}} L_{\text{reg}}$, where L_{data} measures prediction error, L_{phys} penalizes violations of physical relationships, and L_{reg} controls model complexity. Physics-guided graph learning provides a related direction

in which process structure and physical knowledge are incorporated into graph-based soft sensors (Liu et al., 2024).

8. Explainable AI and SHAP-Based Interpretation

8.1. Need for explainability

Interpretability is important when predictions influence process operation, maintenance, optimization, or engineering diagnosis. Explainable AI can help identify which measurements are most influential for a particular prediction and whether the model relies on variables that are consistent with process understanding. Work on explainable AI for fault-classification models in chemical engineering demonstrates how feature-attribution methods can be used to interpret deep-learning outputs (Bhakte et al., 2022).

8.2. SHAP methodology

SHAP is based on Shapley-value concepts from cooperative game theory and assigns contributions to input features relative to a reference prediction. In a catalyst-performance model, SHAP values can be used at both global and local levels. Global analysis can identify variables that consistently influence predictions across a dataset, whereas local explanations can show why a specific operating point received a particular predicted yield or catalyst-performance estimate.

Engineering applications have demonstrated the use of SHAP for interpreting machine-learning predictions. For example, SHAP-based analysis has been used to investigate the interpretability of machine-learning models for fuel higher heating value prediction (Sharma Timilsina et al., 2024). In FCC, analogous analyses could reveal the relative contributions of reactor temperature, catalyst-to-oil ratio, regenerator temperature, feed properties, catalyst activity, and other variables to predicted conversion or product yields.

9. Proposed Comparative Framework

The literature reviewed above suggests that no single model family simultaneously maximizes mechanistic fidelity, predictive flexibility, interpretability, and computational efficiency. A useful framework should therefore compare candidate architectures along several dimensions: prediction accuracy, physical consistency, temporal generalization, robustness to process drift, interpretability, training cost, inference cost, and data requirements.

A practical development strategy begins with a transparent baseline. A conventional data-driven soft sensor establishes the predictive value of available measurements. CNN or temporal architectures can then be introduced to exploit sequential information. Graph-based models can incorporate process topology, while physics-guided losses or mechanistic residual learning can constrain the model. Finally, SHAP and complementary diagnostics can be applied to determine whether the learned relationships remain compatible with process knowledge.

10. Proposed Data and Modeling Pipeline

10.1. Data acquisition and preprocessing

The input dataset should combine routinely measured process variables with catalyst, feed, and laboratory-derived targets. Variables should be synchronized to a common time basis, and laboratory measurements should be aligned with the process residence time and sampling delay.

Missing data should be treated explicitly rather than replaced without considering process context. Robust scaling, outlier detection, and operating-regime identification can be incorporated before model training.

10.2. Feature engineering

Feature engineering should preserve both instantaneous and dynamic information. Candidate engineered features include moving averages, temporal gradients, residence-time-related lagged variables, catalyst circulation indicators, and ratios that have direct process meaning. Feed descriptors and catalyst-state variables should be included when available. Care should be taken to avoid constructing features that inadvertently use future information.

10.3. Model training and validation

A chronological train-validation-test strategy is preferable for dynamic industrial processes because random splitting can place highly correlated observations from the same operating episode into both training and test sets. Independent campaign validation is particularly informative when the objective is deployment under changing catalyst age, feed composition, or operating targets. The evaluation should include both average predictive error and performance under different operating regimes.

11. Evaluation Metrics and Physical Consistency

Predictive accuracy should be reported using several complementary metrics. Mean absolute error is easy to interpret and less sensitive to large errors than squared-error metrics. Root mean square error emphasizes larger deviations, while R^2 provides a measure of explained variance. However, none of these metrics alone indicates whether predictions are physically meaningful.

A physics-guided model should therefore be evaluated using additional consistency metrics. These may include the percentage of predictions violating non-negativity constraints, mass-balance residuals, energy-balance residuals, monotonicity violations where a relationship is known to be monotonic within a defined operating regime, and stability under extrapolation. Such metrics help distinguish a numerically accurate model from one that is also useful for engineering deployment.

12. Computational Trade-Offs and Real-Time Deployment

The principal motivation for replacing or augmenting high-fidelity mechanistic simulation with deep learning is the difference between offline computational cost and online inference cost. A detailed hydrodynamic model may require substantial computational resources, while a trained neural network can often evaluate a prediction rapidly. Nevertheless, the offline generation of simulation data, model training, hyperparameter optimization, and periodic retraining can remain expensive.

A promising architecture is therefore hierarchical. High-fidelity kinetic and hydrodynamic models can be used offline to generate physically informative datasets or constraints, while a compact deep-learning surrogate operates online. The surrogate can be periodically recalibrated using plant measurements. This approach preserves a connection to mechanistic understanding while reducing online computational overhead.

13. Digital-Twin Perspective

A digital twin for an FCC system can be interpreted as a continuously updated computational representation of the physical process. Within the framework proposed here, mechanistic models provide structural knowledge, process data provide observations, and deep learning provides fast nonlinear inference. The result is not necessarily a single monolithic model but a layered system in which different model classes perform complementary functions.

The reactor–regenerator loop is particularly suitable for this concept because process topology, catalyst circulation, thermal coupling, and reaction performance are naturally interconnected. Graph representations can encode this topology, temporal architectures can encode dynamic behavior, and physics-guided terms can restrict predictions to plausible process states. Explainability modules can then provide engineering interpretation of model outputs.

14. Research Gaps and Future Directions

Several research gaps remain before physics-guided deep-learning soft sensors can become routine tools for FCC catalyst-performance prediction. First, high-quality datasets linking catalyst state, feed composition, operating conditions, and laboratory product measurements are required. Second, model validation should explicitly address changes in feedstock and catalyst condition rather than relying only on interpolation within a narrow operating domain. Third, physical constraints should be formulated carefully so that they improve generalization without suppressing legitimate nonlinear process behavior.

Fourth, multiphase hydrodynamics should be integrated more systematically into data-driven models. Existing hydrodynamic studies demonstrate that distributor geometry, baffles, catalyst distribution, and cooler configuration can affect flow and heat transfer (Li et al., 2020a, 2020b; Liu et al., 2022; Oloruntoba et al., 2022; Yang et al., 2021a, 2021b). A future soft sensor could use reduced-order hydrodynamic descriptors derived from detailed simulations rather than attempting to reproduce every flow field directly.

Fifth, explainability should move beyond static feature rankings. Local explanations, temporal attribution, uncertainty estimation, and counterfactual analysis could provide more actionable information. In addition, explanations should be checked against established process knowledge rather than interpreted solely as evidence of causality. SHAP identifies feature contributions to model outputs; it does not by itself establish causal relationships.

Finally, uncertainty quantification should become a standard component of deployment. A model should distinguish between predictions made within a well-represented operating domain and predictions made under conditions that differ substantially from training data. This is particularly important for refinery applications in which feed and catalyst conditions evolve over time.

15. Integrated Proposed Architecture

Based on the reviewed literature, an integrated architecture can be organized into five layers. Layer 1 is the physical-process layer containing feed, reactor, regenerator, catalyst circulation, cooling, and separation variables. Layer 2 is the mechanistic layer containing lumped kinetics, material and energy relationships, and reduced-order hydrodynamic descriptors. Layer 3 is the representation-learning layer containing temporal convolutions, recurrent or attention mechanisms, and graph-based representations. Layer 4 is the physics-guidance layer in which mechanistic relationships contribute constraints or residual-learning targets. Layer 5 is the

interpretation and deployment layer, incorporating SHAP-based explanations, uncertainty indicators, monitoring, and real-time inference.

16. Discussion

The reviewed evidence supports a shift from viewing mechanistic and data-driven models as competing alternatives toward treating them as complementary components. Lumped kinetics provide a compact description of reaction pathways; hydrodynamic models provide insight into gas–solid contacting and catalyst distribution; deep learning provides flexible nonlinear approximation; graph learning encodes plant topology; and XAI provides an interpretive layer. The principal design challenge is deciding how much information should be transferred between these components.

For applications focused primarily on short-term prediction within a stable operating region, a temporal deep-learning soft sensor may provide an efficient solution. When plant-wide interactions are important, graph-based representations become more attractive. When extrapolation and physical consistency are critical, physics-guided learning provides an additional safeguard. When predictions must be reviewed by process engineers, explainability and uncertainty information should be treated as core model outputs rather than optional post-processing.

The choice between FCC and hydrocracking should ultimately depend on the available dataset and the intended prediction target. The reference set supplied for this manuscript provides particularly strong support for FCC kinetics, regeneration, hydrodynamics, and catalyst circulation. Therefore, FCC is the most directly supported primary case for implementation, while the same deep-learning methodology can be adapted to hydrocracking once an appropriately matched hydrocracking reference set and dataset are available.

An important methodological principle is that high predictive accuracy should not be interpreted as evidence that a model has learned causal process mechanisms. Similarly, a physically constrained model is not automatically correct if its physical assumptions are incomplete or incorrectly formulated. Robust development therefore requires parallel evaluation of statistical accuracy, physical consistency, interpretability, computational efficiency, and generalization.

17. Conclusions

This review and framework examined the use of deep learning for predicting FCC or hydrocracking catalyst performance, with FCC serving as the principal case supported by the supplied literature. FCC is a highly coupled catalytic and multiphase system in which reaction kinetics, catalyst deactivation, regeneration, heat transfer, and gas–solid hydrodynamics jointly determine process performance. Lumped kinetic models offer tractable mechanistic representations, while integrated reactor–regenerator and hydrodynamic models provide additional physical detail (Chen et al., 2020; Du et al., 2019; Olafadehan et al., 2019; Wang, 2020).

Deep-learning soft sensors provide a complementary route for rapid prediction of difficult-to-measure process variables. Temporal and multichannel convolutional architectures can exploit dynamic patterns, while spatiotemporal and graph-based models can represent local structure and plant-wide relationships (Wang et al., 2024; Wu et al., 2024; Yuan et al., 2020). Physics-guided learning extends these capabilities by incorporating process knowledge into model training or

architecture, with recent chemical-process research demonstrating the potential of physics-guided graph learning for soft sensing (Liu et al., 2024).

Explainability is essential for practical adoption. SHAP-based analysis can expose the contribution of individual process variables to predictions and help engineers examine whether the model uses plausible information (Bhakte et al., 2022; Sharma Timilsina et al., 2024). However, such explanations should be interpreted as model attribution rather than direct causal evidence.

Overall, the proposed framework supports the development of physics-informed, spatiotemporal, graph-based, and explainable soft sensors that operate as fast surrogates for more computationally demanding mechanistic models. The most promising research direction is not a universal replacement of first-principles simulation, but a layered digital-twin architecture in which mechanistic models, hydrodynamic information, deep learning, physical constraints, and explainability are integrated according to the requirements of the target refinery application. Future experimental and industrial studies should focus on independent validation under feed and catalyst changes, uncertainty quantification, physical-consistency testing, and real-time deployment.

Acknowledgment

The author acknowledges the use of artificial intelligence (ChatGPT version 5.6) as an assistive tool to refine the grammar, syntax, and clarity of this manuscript. The author assumes full responsibility for the final content, arguments, and accuracy of the presented work.

References

- Bhakte, A., Pakkiriswamy, V., & Srinivasan, R. (2022). An explainable artificial intelligence based approach for interpretation of fault classification results from deep neural networks. *Chemical Engineering Science*, 250, 117373.
- Carter, M. A., Camille, A. K., Fewel, K., Ciobotaru, R., & Diwanji, B. (2020). Catalyst heat removal cluster and tube design (U.S. Patent No. 10,782,075 B2). U.S. Patent and Trademark Office.
- Chen, Y., Wang, W., Wang, Z., Hou, K., Ouyang, F., & Li, D. (2020). A 12-lump kinetic model for heavy oil fluid catalytic cracking for cleaning gasoline and enhancing light olefins yield. *Petroleum Science and Technology*, 38(10), 912–921.
- Davydov, L., Sandacz, M. S., Palmas, P., Fei, Z., & Lippmann, M. (2020). FCC counter-current regenerator with a regenerator riser (U.S. Patent No. 10,751,684 B2). U.S. Patent and Trademark Office.
- Du, Y., Sun, L., Berrouk, A. S., Zhang, C., Chen, X., Fang, D., & Ren, W. (2019). Novel integrated reactor-regenerator model for the fluidized catalytic cracking unit based on an equivalent reactor network. *Energy & Fuels*, 33(8), 7265–7275.
- Ferreira, J., Pedemonte, M., & Torres, A. I. (2022). Development of a machine learning-based soft sensor for an oil refinery's distillation column. *Computers & Chemical Engineering*, 161, 107756.
- Fisher, O. J., Watson, N. J., Escrig, J. E., Witt, R., Porcu, L., Bacon, D., Rigley, M., & Gomes, R. L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers & Chemical Engineering*, 140, 106881.
- Hsu, C. S., & Robinson, P. R. (2019a). Cracking. In *Petroleum science and technology* (pp. 211–244). Springer.
- Hsu, C. S., & Robinson, P. R. (2019b). Gasoline production. In *Petroleum science and technology* (pp. 189–210). Springer.
- Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795–814.
- Lesage, R., & Simon, H. (2020). Anchoring structure for an anti-erosion coating, in particular for protecting a wall of an FCC unit (U.S. Patent No. 10,799,846). U.S. Patent and Trademark Office.
- Li, J., Yao, X., Liu, L., & Lu, C. (2020a). Investigation on distribution of particles in inlet region of an FCC external catalyst cooler with different inlet structures. *Powder Technology*, 362, 267–277.
- Li, J., Yao, X., Liu, L., Zhong, X., & Lu, C. (2020b). Heat transfer characterization and improvement in an external catalyst cooler fluidized bed. *Particuology*, 56, 103–112.
- Liu, D., Zhang, Y., Zhang, S., & Oloruntoba, A. (2022). Effect of structure parameters on forces acting on the baffles used in gas-solids fluidized beds. *Particuology*, 61, 111–119.
- Liu, Y., Jia, M., Xu, D., Yang, T., & Yao, Y. (2024). Physics-guided graph learning soft sensor for chemical processes. *Chemometrics and Intelligent Laboratory Systems*, 249, 105131.
- Liu, Y., Liu, W., Lyu, Y., Liu, X., Zhang, J., Gu, Y., Huang, J., Mintova, S., & Yan, Z. (2020). Intra-crystalline mesoporous SAPO-11 prepared by a grinding synthesis method as FCC promoters to increase iso-paraffin of gasoline. *Microporous and Mesoporous Materials*, 305, 110320.
- Nnabalu, C. R., Falcone, G., & Bortone, I. (2019). The role of fluid catalytic cracking in process optimisation for petroleum refineries. *International Journal of Chemical and Molecular Engineering*, 13(6), 370–376.
- Olafadehan, O., Daramola, O., Sunmola, O., & Abatan, G. (2019). Modelling and simulation of riser reactor of a commercial fluid catalytic cracking unit using 6-lump kinetics of vacuum gas oil. *Petroleum & Petrochemical Engineering Journal*, 3(1), 000194.
- Oloruntoba, A., Zhang, Y., & Hsu, C. S. (2022). State-of-the-art review of fluid catalytic cracking (FCC) catalyst regeneration intensification technologies. *Energies*, 15(6), 2061.

- Oloruntoba, A., Zhang, Y., & Xiao, H. (2022). Study on effect of gas distributor in fluidized bed reactors by hydrodynamics-reaction-coupled simulations. *Chemical Engineering Research and Design*, 177, 431–447.
- Perera, Y. S., Ratnaweera, D. A. A. C., Dasanayaka, C. H., & Abeykoon, C. (2023). The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review. *Engineering Applications of Artificial Intelligence*, 121, 105988.
- Qin, R., & Zhao, J. (2022). Adaptive multiscale convolutional neural network model for chemical process fault diagnosis. *Chinese Journal of Chemical Engineering*, 50, 398–411.
- Sabahi, A., Loebel, A. J., Galvada, S., Francis, J. A., & Iyyamperumal, E. (2019). FCC catalyst with more than one silica, its preparation and use (U.S. Patent Application No. 16/322,779). U.S. Patent and Trademark Office.
- Sadeghbeigi, R. (2020). *Fluid catalytic cracking handbook: An expert guide to the practical operation, design, and optimization of FCC units* (4th ed.). Butterworth-Heinemann.
- Sexton, J. A. (2019). FCC cyclone using acoustic detectors (U.S. Patent No. 10,239,034). U.S. Patent and Trademark Office.
- Sharma Timilsina, M., Sen, S., Uprety, B., Patel, V. B., Sharma, P., & Sheth, P. N. (2024). Prediction of HHV of fuel by machine learning algorithm: Interpretability analysis using Shapley additive explanations (SHAP). *Fuel*, 357, 129573.
- Speight, J. (2019). Catalytic cracking processes. In *Heavy oil recovery and upgrading* (pp. 357–421). Gulf Professional Publishing.
- Sun, Q., & Ge, Z. (2021). A survey on deep learning for data-driven soft sensors. *IEEE Transactions on Industrial Informatics*, 17(9), 5853–5866.
- Wang, J. (2020). Continuum theory for dense gas-solid flow: A state-of-the-art review. *Chemical Engineering Science*, 215, 115428.
- Wang, X., Qi, X., & Zhang, Y. (2024). Deep learning with local spatiotemporal structure preserving for soft sensor development of complex industrial processes. *Applied Soft Computing*, 164, 111974.
- Wu, W., Song, C., Zhao, J., & Wang, G. (2024). Knowledge-enhanced distributed graph autoencoder for multiunit industrial plant-wide process monitoring. *IEEE Transactions on Industrial Informatics*, 20(2), 1871–1883.
- Yan, C.-Y., Sun, G.-G., Chen, J.-Y., Wei, Y.-D., & Lu, C.-X. (2020). Experimental study on the particle flow patterns in a cyclone dipleg with a trickle valve. *Petroleum Science and Technology*, 17(11), 822–837.
- Yang, X., Liu, G., Li, Y., Zhang, L., Wang, X., & Liu, Y. (2019). Novel Pt–Ni bimetallic catalysts Pt (Ni)–LaFeO₃/SiO₂ via lattice atomic-confined reduction for highly efficient isobutane dehydrogenation. *Transactions of Tianjin University*, 25(3), 245–257.
- Yang, Z., Zhang, Y., Liu, T., & Oloruntoba, A. (2021a). MP-PIC simulation of the effects of spent catalyst distribution and horizontal baffle in an industrial FCC regenerator. Part II: Effects on regenerator performance. *Chemical Engineering Journal Advances*, 6, 100095.
- Yang, Z., Zhang, Y., Oloruntoba, A., & Yue, J. (2021b). MP-PIC simulation of the effects of spent catalyst distribution and horizontal baffle in an industrial FCC regenerator. *Chemical Engineering Journal*, 412, 128634.
- Yang, Z., Zhang, Y., & Zhang, H. (2019). CPFD simulation on effects of louver baffles in a two-dimensional fluidized bed of Geldart A particles. *Advanced Powder Technology*, 30(11), 2712–2725.
- Yuan, X., Qi, S., Shardt, Y. A. W., Wang, Y., Yang, C., & Gui, W. (2020). Soft sensor model for dynamic processes based on multichannel convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, 203, 104050.

- Zhang, D., Dai, Y., Chen, K., Guo, A., Liu, H., & Wang, Z. (2021). Agglomeration of solid particles in fluidized catalytic cracking slurry oil: Particle separation by the oil–water interface and particle composition analysis. *Energy & Fuels*, 35(4), 3219–3227.
- Zhang, L., Pilot, I. A., Su, Y.-Q., Liu, J.-X., & Hensen, E. J. (2019). Understanding the impact of defects on catalytic CO oxidation of LaFeO₃-supported Rh, Pd, and Pt single-atom catalysts. *The Journal of Physical Chemistry C*, 123(12), 7290–7298.