Front. Chem. Sci. Eng. 2025, 19(7): 57
https://doi.org/10.1007/s11705-025-2562-1

REVIEW ARTICLE

Multi-scale revolution of artificial intelligence in chemical industry

Ying Li[1], Quanhu Sun[1], Zutao Zhu[1], Huaqiang Wen[1], Saimeng Jin[1], Xiangping Zhang[2], Zhigang Lei[3], Weifeng Shen (✉)[1]

1 School of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China
2 College of Chemical Engineering and Environment, China University of Petroleum-Beijing, Beijing 102249, China
3 School of Chemistry and Chemical Engineering, Shihezi University, Shihezi 832003, China

Abstract With the advent of the fourth technological revolution, the new generation of artificial intelligence (AI) has imparted new significance and opportunities to the modeling of momentum, heat, and mass transfer, as well as chemical reaction processes with the realm of chemical engineering. AI techniques are being widely employed in the chemical industry and are constantly evolving to offer more effective solutions for tackling practical challenges. This review delves the transformation of the chemical industry from traditional digital simulations to advanced AI-based approaches, targeting high efficiency and low carbon emissions across the scale from molecules to factories. Particular emphasis is mainly placed on the research carried out within the research group of Weifeng Shen. At the molecular level, the intelligent capture of molecular characteristics and the precise determination of structure-property relationships have reached a mature stage. Furthermore, multifunction-driven reverse molecular design for solvents, reaction reagents, and other substances has been accomplished through AI-based high-throughput screening and generative models. To improve the safety, environmental friendliness, and carbon reduction performance of chemical separation processes, a series of innovative reinforcement strategies have been put forward, with a primary focus on the systematic optimization of solvent design. On the process scale of actual production, it frequently occurs that the constructed mechanism model fails to align with the actual system behavior, thereby restricting the industrial application of the model. To solve this issue, mechanism-data hybrid-driven frameworks have been successfully developed, leveraging AI-enhanced prediction, diagnosis, optimization, and control for complex separation systems in practice. Finally, as a bridge connecting big data intelligent technology and actual industrial processes, dynamic digital twin modeling is discussed for its potential to boost efficiency and sustainability in the chemical industry.

Received December 18, 2024; accepted April 7, 2025; online May 23, 2025

E-mail: shenweifeng@cqu.edu.cn

1 Introduction

Chemical system engineering is a comprehensive discipline that encompasses automatic control [1], product design [2], computer science [3], and related fields. As a cornerstone sector, the chemical industry is characterized by high resource and energy consumption with significant environmental impacts [4]. The production processes within the chemical industry involve a complex chain from raw material extraction to final product manufacturing, with frequent energy and material transfers often resulting in inefficiencies and waste. In response to these challenges, the industry is progressively shifting toward economically and environmentally sustainable production methods [5]. The demand for clean, low-carbon, and high-efficiency production technologies continues to drive the greening of chemical processes [6], which requires comprehensive multi-scale optimization from the molecular to the factory scale.

At the molecular scale, the intensification of chemical industrial processes includes the selection of solvents [7], catalysts [8], and gas absorbents [9], which significantly influence overall process improvement. The development of high-performance molecules aims to meet criteria like green efficiency and safety. Traditional research methods relying on trial-and-error approaches are resource-intensive and time-consuming [10]. Hence, the efficient and rapid design of high-performance molecules is crucial for industrial green transformation.

At the factory scale, the optimization objectives for

chemical industrial processes include reducing energy consumption, minimizing environmental pollution, maximizing resource utilization, and enhancing process safety [11]. Tackling the trade-offs between multiple constraints in chemical industrial process optimization has long been a focus for researchers [12]. Although recent studies have developed optimization methods, traditional ones often fall short in effectiveness. As optimization problems become more complex, conventional mathematical optimization tools frequently become inadequate or prone to local optima. Furthermore, accurate mathematical modeling of chemical processes is crucial for plant-level decision-making, and efficient data modeling frameworks are needed given the complexity of chemical systems and vast amounts of data.

Artificial intelligence (AI) technology, with its outstanding data processing capabilities, efficient pattern recognition, and strong scalability, offers new perspectives for solving complex scientific problems. As industrial scales expand and the volume of data grows rapidly alongside state-of-the-art information technologies, data-driven methods hold promising prospects for intelligent manufacturing [13] and facilitate the transition from automated [14] to intelligent production [15]. From the molecular level, AI can learn chemical knowledge and effectively design new high-property molecules that can be potentially synthesized [16]. AI can also make more choices based on specific requirements and consider more factors in new research domains. By combining AI with human intelligence, theoretical references and guidance can be provided, and both scientifically validated and prospectively successful routines can be unearthed with relatively few resources. From the factory level, AI helps optimize production flows and conditions by analyzing production data, reducing the environmental costs from by-products and manufacturing waste. Additionally, AI enables timely troubleshooting and alarms through real-time monitoring technology [17], enhancing safety and operational stability in industrial processes. Leveraging AI makes multi-objective optimization achievable [18], allowing for the consideration of numerous variables to reach globally optimal conclusions. Compared to traditional mathematical and statistical tools, the exceptional processing capacity of AI can handle large amount of variable data. In this era, the design, optimization, and intensification of chemical processes will undoubtedly benefit from the significant role of AI [19].

This review focuses on the current progress in the comprehensive optimization of chemical industrial processes from molecular to factory scales based on the new generation of AI techniques. At the molecular scale, innovative solutions for quantitative structure-property relationship (QSPR) modeling based on AI are elaborated, including efficient feature extraction techniques for molecular structures, improvements in building AI model algorithm architectures by incorporating chemical engineering knowledge to enhance model performance. At the process scale, multi-objective molecular generation techniques utilizing the synergy between QSPR models and AI are introduced, achieving collaborative optimization of high-performance molecules and process design. At the factory scale, hybrid mechanistic modeling optimization methods for process systems based on AI are discussed, aimed at improving energy utilization efficiency and reducing carbon emissions. The real-time dynamic quality control of chemical production processes using digital twin technology is presented. In light of existing research achievements, it is believed that AI-based multiscale optimization paradigms in the chemical industry have substantial potential for achieving cleaning, energy-efficient, and carbon-reducing production processes [20]. The maturity of dynamic digital twin modeling technology through the combination of AI-related advanced technologies will promote real-time monitoring and optimization of chemical processes, enable long-term prediction of processes, and provide more accurate decision support [21]. The dynamic digital twin modeling technology will play an increasingly important role in various intelligent production systems, empowering enterprises with digital transformation and intelligent operation.

2 AI-based prediction of physicochemical properties at the molecular scale

To build a predictive model for molecular properties, it is vital to capture the physical, chemical, and structural information of the molecule and translate it into a mathematical format for effective model input [22]. QSPR utilizes mathematical and statistical tools to uncover quantitative relationships between chemical kinetics of compounds and molecular structures, and physical and chemical properties [23]. The QSPR modeling process for property prediction generally includes several steps (as depicted in Fig. 1): (1) data collection and preprocessing, (2) structure cognition and feature extraction, (3) model establishment, (4) performance evaluation, and (5) analysis and reinforcement. The detailed key steps will be presented in the following subsections.

2.1 Structural representation and feature extraction methods

Structure cognition and feature extraction are fundamental components of these processes, choosing appropriate approaches can significantly enhance effectiveness. Researchers have made significant progress in both molecular structure characterization [24] and feature extraction techniques [25]. Commonly used methods

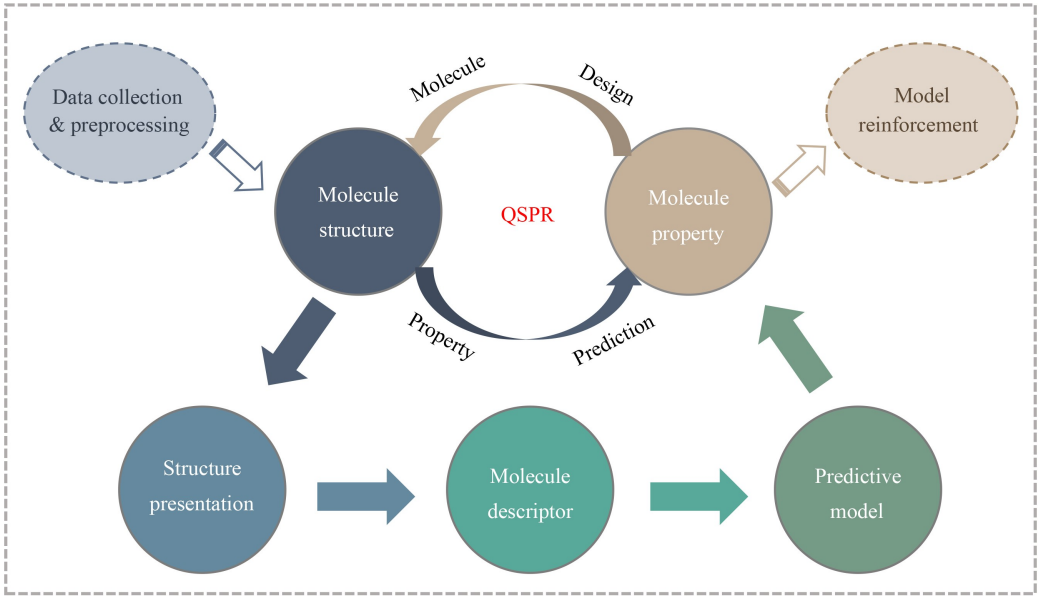


Fig. 1 The process of molecule property prediction via AI-based QSPR modeling.

range from the earlier proposed group contribution method (GCM) to current molecular descriptors, molecular fingerprinting, and other novel methods.

2.1.1 GCM

GMC has been universally used for calculating and predicting variable properties, including thermodynamic behavior, which is helpful for computer-aided molecule design [26]. It is considered a one-dimensional representation with limitations in expressing connectivity and interrelationships between substructures, based on the theory that molecular properties are additive and individual structural units contribute equally in molecules [27].

Hu et al. [28] conducted a comprehensive screening of organic solvents and ionic liquids to identify optimal entrainers for separating aqueous systems from *tert*-butanol. The approach considered both the toxicity and physical properties of the candidate solvents while using GCM to evaluate the melting points and thermal decomposition temperatures. The study offers innovative perspectives for solvent design and can guide the selection of azeotropes in other separation processes.

Although GCM is reliable for simple models, it has drawbacks in practical applications, like calculation errors and difficulties with complicated structures. However, constructing complex nonlinear GCM models can improve accuracy and efficiency in estimating properties.

2.1.2 Molecular descriptors and fingerprints

In cheminformatics, molecular descriptors play a key role in the quantitative expression of molecular characteristics. Molecular descriptors can be divided into experimental and theoretical descriptors [29]. Experimental descriptors, like viscosity, density, and polarity, are derived from empirical measurements, while theoretical descriptors, such as molecular fingerprints, molecular diagrams, and molecular structural fragments, are calculated using computational methods. These two types of descriptors are complementary, with experimental ones validating or calibrating theoretical models and theoretical ones enabling property prediction for molecules lacking experimental data. Their classification guides the selection of appropriate descriptors for specific tasks [30]. Molecular descriptors quantify molecular characteristics and deepen the computational understanding of molecular structures and activities [31]. The typical molecular descriptors encompass the atomic charge, hydrogen bonding effects, molecular volumes, surface areas, along with a diverse range of derived descriptors [32].

Molecular fingerprints have long been employed in drug discovery [33] and virtual screening [34]. They function as property profiles for molecules and are usually represented as binary strings that signify specific molecular substructures. The type of molecular fingerprints depends on the methodology used to define these substructures. Examples include substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints. Zhu et al. [35] proposed a novel adaptive architecture for predicting the hazardous properties of industrial chemicals, focusing on various toxicological endpoints related to acute oral toxicity. The authors' research group integrated multiple molecular fingerprints into a multivariate model, which helped overcome the limitations inherent in individual fingerprints. The framework addressed descriptor selection and model interpretability by providing a unique method for sequence-based molecular characterization.

Additionally, it introduced an adaptive modeling architecture and a new feature perturbation method for data evaluation, thereby enhancing the understanding of toxicity mechanisms.

2.1.3 Novel feature extraction methods

Over the past few decades, an ever-increasing number of advanced feature extraction methods have been developed [36]. Building on these novel approaches for molecular structure characterization, several promising advancements have emerged. Recent molecular feature extraction methods include graph, natural language processing (NLP), molecular image, three-dimensional (3D) structure representation [37], and computer vision [38]. Molecular graph-based representation learning techniques have demonstrated remarkable competitiveness. Graphs are one of the most intuitive protocols to represent molecular structures [39]. The molecular graph serves as an effective two-dimensional (2D) molecular descriptor, transforming molecules into matrices that represent topological features, where atoms and bonds are regarded as nodes and edges, respectively [40].

To deal with the problem of accurately predicting the infinite dilution activity coefficient (γ^{∞}), the research group of Shen [41] conducted a comprehensive study on an accurate predictive model based on a graph-learning architecture. The model has four main stages: (1) molecules are represented as undirected graphs; (2) a message-passing neural network (MPNN) is applied to extract intra-molecular features, during the process, the surrounded chemical spatial information clusters on atoms; (3) an interactive attention model is then developed to capture the hydrogen-bonding interactions between solute and solvent; (4) the final features generated by graph-learning incorporate comprehensive information on both intra- and intermolecular features as well as temperature-dependent parameters. The presented model exhibits superior performance in terms of accuracy and reliability and holds a promising application prospect in the area of green solvent screening and actual separation processes.

The molecular graph-based hybrid representation methods mainly concentrate on 2D-based features at the atom, bond, or molecule level, often overlooking the 3D spatial structure information that is crucial for explaining molecular properties. These methods may also fail to incorporate relevant chemical knowledge and molecular stereo-structural information, which can significantly boost model interpretability. Consequently, Zhang et al. [42] combined a directed MPNN, chemically synthesizable fragment features, and 3D spatial structure information to construct a 3D multi-hierarchical representation-based deep neural network framework aimed at predicting environmental, health, and safety (EH&S) properties. Such a representation enables the simultaneous extraction of the local, global, and spatial geometric information from molecules.

NLP is a set of data-driven computational techniques to learn, understand, and generate human language content [43]. The traditional machine learning (ML) approaches based on NLP typically relied on time-consuming and usually incomplete manually crafted features [44]. With the progress in computational power and techniques, NLP has demonstrated remarkable capabilities in computational tasks. Motivated by successful cases in AI-assisted NLP, Su et al. [45] developed a data preparation strategy that encodes molecules with canonical molecular signatures and utilizes an embedding algorithm to vectorize bod-substrings. Among these, tree-structured long short-term memory (LSTM) imitated the tree structures of canonical signatures and output a feature vector that was used to correlate properties within a back-propagation neural network. The innovative aspect of the study lies in the fact that molecules do not need to be represented as images or expressed in a linear language. Meanwhile, the feature extraction method can automatically learn QSPRs and consider a greater number of materials to enhance predictive ability.

The previous work [46] also reported a deep pyramid convolutional neural network architecture via NLP. The research discovered that the simplified molecular input line entry system (SMILES) is essentially a language from which chemical information can potentially be extracted using NLP. It is a string format that employs ASCII characters to represent molecules and chemical reactions. In the new QSPR modeling paradigm, each SMILES notation was interpreted as a sentence, and the molecular eigenvectors output by the model can be reversed back to SMILES. Subsequently, based on two strategies (rule-based and frequency-based), two tokenizers were applied to convert SMILES strings into indexical arrays. The work on array conversion, embedding, and featurization, and reliable predictions of molecular properties was then carried out. It is especially worth noting that the established model can be easily integrated with other SMILES-based models and enables inverse molecule design for new product discovery.

2.1.4 Feature selection and optimization strategy

Feature selection aims to filter out redundant features to ensure model performance [47]. The process enhances the accuracy and effectiveness of the model [48]. In the study conducted by Yang et al. [49], three ML models were used to predict the solubility of CO_2 in ionic liquids, incorporating three different molecular descriptors: GCM, molecular structure descriptors (MSD), and mixed GCM-MSD. Model evaluation was performed by mean absolute error, coefficient of determination (R^2), and mean relative error. The findings of the study provide valuable insights into the potential molecular characteristics that influence

the solubility of carbon dioxide in ionic liquids and lay a foundation for future research. The results revealed that the predictive performance of the models is more significantly influenced by descriptors with richer feature information, with the CatBoost-group contribution (GC)-MSD model standing out among them.

The octanol-water partition coefficient ($\log K_{ow}$) is a commonly used indicator for evaluating the environmental effects of chemicals [50]. Huang et al. [51] proposed an adaptive dual-optimization feature screening strategy utilizing RDKit as molecular features. The study integrated genetic algorithms and random forests to eliminate redundancy and identify the optimal set of molecular descriptors. The model was developed combining a back-propagation neural network and Bayesian optimization, enabling automatic integration to evaluate learning validity. The model shows good accuracy and high reliability in predicting the $\log K_{ow}$ of organic molecules, thereby contributing to the discovery of green solvents.

The traditional GCM has limitations when it comes to accurately predicting properties in practical applications. To tackle this issue, Shen's research team [52] proposed a novel strategy for extracting molecular features that are both interpretable and effective in distinguishing isomers. The approach can rapidly identify molecular structures and extract molecular features from SMILES representations. Each molecular feature corresponds to a substructure composed of a non-hydrogen atom, their connected hydrogen atoms, and associated chemical bonds. Different types of chemical information, including the type of the non-hydrogen atom, the number of hydrogen atoms, formal charge, and bond types between the substructure and its neighbors were considered. The frequency of these molecular features was then utilized as structural descriptors, enabling ML algorithms to establish correlations between molecular structures and properties. The comprehensive chemical information encoded into molecular features allows the resulting models to effectively distinguish isomers. Notably, fewer molecular features are required to build accurate predictive models using the proposed feature extraction strategy.

2.2 AI-based QSPR modeling

In the last decades, using QSPR models to solve chemical problems has become increasingly prevalent. The QSPR model undergoes several stages, which can be roughly categorized into three types: (1) classical statistical methods [53], (2) basic ML models, and (3) deep learning (DL) frameworks. Traditional methods have not been phased out; rather, the integration of classical statistical tools with modern advanced technologies like ML and DL has emerged as a major trend. In recent years, AI has almost taken the lead in various fields, including QSPR, due to its rapid growth and development [54].

ML is a data analysis method capable of uncovering underlying patterns and rules from given cases. It is primarily divided into three classes: supervised learning, unsupervised learning, and reinforcement learning [55]. Basic ML models include support vector machine [56], decision tree [57], random forest [58], and neural networks [59]. Different ML models constructed using various algorithms exhibit significant discrepancies in performance. Therefore, employing appropriate algorithms can lead to more accurate results. Although ML performs well, it still has limitations when dealing with more complex problems.

DL-based frameworks excel at automatically identifying patterns in complex nonlinear data sets, effectively serving as automated feature engineering. The application of DL technology is gradually spreading to an increasing number of fields, attracting growing interest from researchers due to advantages in data processing efficiency, scalability, robust performance, and support for multi-task learning and other factors. The application of DL in chemistry is extensive and has greatly accelerated research progress and led to significant advancements [60].

Inspired by NLP, Zhang et al. [61] proposed a feedforward neural network based on MPNNs for the rapid and accurate prediction of surface charge density profiles and cavity volumes in conductor-like screening models (COSMO) segment activity coefficient. The approach tackled the issue that the original GC-COSMO was unable to distinguish between isomers. By employing a hybrid molecular representation, the novel model effectively differentiates between *cis*/*trans* and structural isomers. Additionally, it enables the rapid calculation of physicochemical properties, facilitating high-throughput screening of green solvents. In future research, the study will continue to integrate the COSMO-realistic solvation database to expand cavity volume and σ-profile data sets, enhancing the diversity of molecular types and training more generalized predictive models.

Wang et al. [62] developed a QSPR model employing DL techniques to predict $\log K_{ow}$ for organic compounds accurately and reliably. The model provides valuable environmental insights that can guide the selection and development of key chemicals such as green solvents. Canonical molecular signatures, derived from canonical molecular graphs, were utilized as input parameters by mapping them onto tree-structured LSTM networks. By integrating these molecular signatures with the tree-structured LSTM, the model minimizes human intervention in molecular feature selection and enables automatic extraction of molecular features. The deep-learning neural network not only predicts the $\log K_{ow}$ to assess the lipophilicity of organic chemicals but also has the potential for broader applications in evaluating other environmental properties to promote the development of green chemistry.

To date, QSPR has attracted increasing attention due to its excellent predictive capability. Compared with first-principles calculations, DL-based QSPR models eliminate the need for complex mechanistic analysis during the modeling process. This not only saves considerable computational resources but also delivers satisfactory accuracy. As a result, more researchers are seeking an effective QSPR model to solve existing or potential chemical problems. This trend is expected to drive advancements in molecular discovery and intelligent chemical process optimization.

2.3 Enhancement strategies for DL modeling based on chemical knowledge

Despite the availability of effective molecular structure characterization methods, QSPR modeling strategies, and ML models, the actual modeling performance is influenced by numerous objective factors. These include factors such as the poor quality of training samples, the complexity of structure-property relationships that make fitting difficult, and the limitations of certain modeling techniques that can lead to challenges like convergence issues, overfitting, high uncertainty, or poor interpretability. Consequently, model enhancement has increasingly become a focus of research, aiming to maximize model performance under existing training conditions. The following will briefly discuss several model enhancement strategies.

2.3.1 Model performance enhancement through few-shot learning and hyperparameter optimization

ML is inherently data-driven; however, the availability of adequate data poses difficulties in practical applications [63]. Several challenges remain, including limited storage capacity, a scarcity of labeled samples, model stability, the quality of generated samples, and an excessive number of parameters, etc. [64]. There is an urgent need to address a learning paradigm known as few-shot learning [65]. The few-shot phenomenon observed in training sets of molecular structures and properties is primarily due to the limitations and sparsity of the chemical spaces relevant to the samples [66].

Through the correlation analysis between QSPR and chemical space clustering, Wen et al. [67] found the few-shot phenomenon and expanded its definition to describe it as the "uneven distribution of training sample quality" within molecular chemical space. The work highlights the impact of quality disparities among samples across different chemical subspaces on model training. Besides, Wen et al. [68] put forward a more general model enhancement strategy for few-shot learning, leveraging a chemical space deconstruction model for sample augmentation in local regions and employing a dynamic ensemble strategy to reinforce the built intelligent model.

Hyperparameter optimization also significantly improves model performance [69], the Hyperopt library with Bayesian optimization and 5-fold cross-validation was used to identify optimal hyperparameters for various ML algorithms [70]. A rank normalized scores approach was also employed to assess whether the ML modules perform better, confirming that while model performance can be improved, there is no universally superior ML algorithm.

2.3.2 Model reliability enhancement through uncertainty analysis

Uncertainty reflects the degree of dispersion of random variables and is typically classified into two categories: aleatory uncertainty and epistemic uncertainty [71]. A key objective of uncertainty analysis is to quantify the outputs associated with uncertainty. The calibration methods currently in use are generally categorized into confidence-based and error-based methods. Commonly used methods of uncertainty analysis include Monte Carlo dropout, model ensemble, and evidential DL (Fig. 2). Some models with simple structures, such as linear regression and Gaussian process regression, can directly estimate model uncertainty, but this capability does not extend to more complex models.

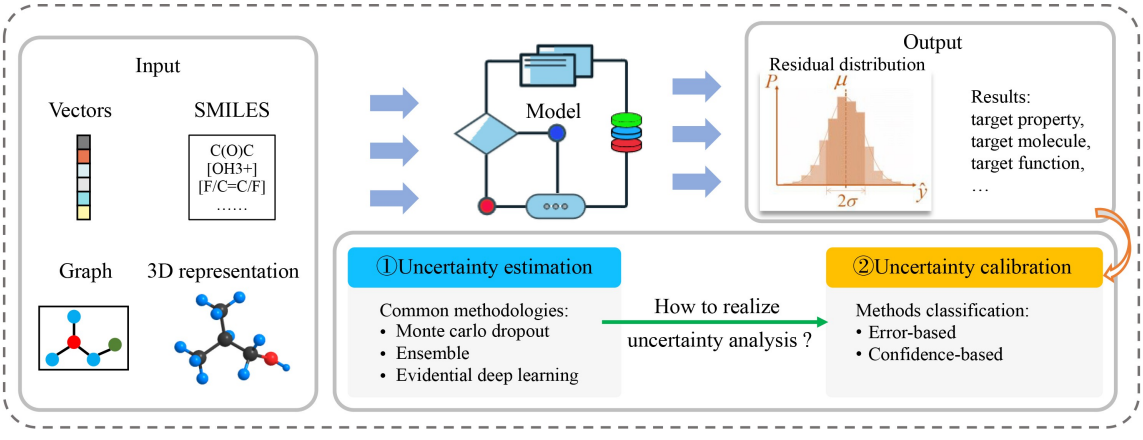


Fig. 2 General workflow of model uncertainty analysis.

The MC-dropout technique has been proven effective by researchers. Wen et al. [67] proposed an approach that is extensively tested by randomly disconnecting node connections in the trained deep neural network, showcasing its robustness and reliability. Models constructed using this approach will offer indicators of predictive uncertainty, assess the reliability of the trained model predictions on molecular data, and help mitigate potential issues in high-risk applications. The specific details of the uncertainty analysis in the study are as follows: (1) for the individual compounds, calculating absolute error and absolute relative error to evaluate predictive ability of the model; (2) for data sets, calculating indicators such as average absolute error, average absolute relative error, root mean square error, and the R^2 to thoroughly assess the overall predictive performance of the model. Through these evaluation metrics, a better understanding of the predictive capabilities of the model and its reliability in practical applications can be attained.

2.3.3 Model interpretability enhancement for uncovering scientific principles

Due to the inherent complexity of DL models, their internal mechanisms are often difficult for humans to comprehend. However, understanding these mechanisms is crucial for unveiling potential scientific principles. Therefore, it is essential to employ interpretability approaches to explain the workings of DL models [72]. Improving model interpretability not only facilitates the discovery of new scientific insights but also helps to explain sudden drops in performance under specific conditions [73]. Significant research has already been conducted in the area of model interpretability analysis, leading to notable advancements in the field [74].

Zhang et al. [75] identified the molecular substructures that contribute most significantly to the target molecule using a Monte Carlo tree search approach. The Monte Carlo algorithm facilitates molecular design through steps including generation, optimization, and property evaluation. The algorithm utilizes target properties and descriptors, employing strategies similar to genetic algorithms to generate molecules that fulfill specific property requirements. Optimization was performed through mutations within an initial pool of molecules. An encoder and specific rules were utilized to avoid generating incorrect SMILES representations. Furthermore, the authors established different ranges of atom counts for substructures based on specific tasks, thereby optimizing the use of computational resources.

In the study previously mentioned [42], the research primarily made use of the breaking of retrosynthetically interesting chemical substructures (BRICS) intelligent cutting algorithm to break down molecules into chemically synthesizable substructures. Through quantitative visualization, it is able to identify the substructures that most significantly contribute to desired properties, guiding molecular design and optimization. The study further demonstrated that the representation achieved by the directed MPNN, which integrated BRICS fragments and 3D information within hierarchical representations, served to enhance the predictive ability for lipophilicity. Moreover, molecular substructures that contribute to sustainability, especially in relation to environmental health and safety aspects, were clustered and analyzed. This analysis provides valuable information that can inform the design of green solvent molecules. The interpretability analysis carried out on these substructures offers effective guidance for the structural design process.

In the study that was previously referred to [41], a novel solute-solvent interactive attention module has been innovated and integrated into the molecular graph-learning architecture. The interactive attention module empowers the model to effectively capture intermolecular hydrogen-bonding interactions. It was specifically designed with the intention of learning pairwise interactions between functional substructures. Additionally, attention coefficients were employed to quantify the significance of these hydrogen-bonding interactions. The developed attention module used the extracted atom-centered substructure information to recognize the critical functional parts within the interactions and then assigned the highest weights to these parts. Subsequently, a readout function known as Set2Set was used to gather the graph-level features for both the solute and the solvent during the readout phase.

Yang et al. [76] conducted a study in which they compared three ML methods, namely feedforward neural networks, extreme gradient boosting, and random forest, for predicting the K_{ow}. The models were rigorously evaluated based on a variety of metrics, including R^2, mean absolute error, root mean square error, and mean relative error. Furthermore, shapeable additive explanations were utilized to enhance the interpretability of the models. The research not only establishes a new benchmark for the accuracy of K_{ow} prediction and addresses the limitations associated with the experimental techniques used for measuring K_{ow}, but it also enhances the interpretability of the QSPR model. Potential future applications of the research include guiding the generation of molecule models.

2.3.4 Multi-task DL for multi-objective QSPR modeling

Previous studies on DL-based QSPR modeling have typically focused on a single property prediction task. However, in practical applications, the performance evaluation of molecules often involves multiple properties [77]. This limitation has led to the emergence of the concept of multi-task DL, which aims to meet the requirement of simultaneously predicting multiple molecular properties. Multi-task DL is a form of

inductive transfer that improves generalization by handling multiple related tasks through the simultaneous optimization of loss functions [78]. The authors proposed a multi-task neural network DL framework (MDNN) [79]. The MDNN can effectively utilize limited experimental data sets to automatically capture relevant molecular features through multiple feedforward neural networks. It can also discern interactions between various attributes via two training strategies: joint training and alternating training. Moreover, it associates different target properties by vectorizing molecular structures through an LSTM network, thereby capturing the generalizability across all tasks without the need for descriptor selection and maximizing the utility of the available experimental data. The study also involves outlier detection and proposes a program that combines the empirical cumulative distribution function of residuals with an analysis of deviations to evaluate the predictive performance of multitask DL models. Partial least squares regression was utilized as a comparative baseline to emphasize the superior performance and scalable data processing capabilities of the proposed MDNN approach. Furthermore, a method based on principal component analysis was introduced to analyze the applicability of the model. Future studies are expected to focus on developing multi-objective molecular generation models that can produce clean, low-carbon solvents with EH&S properties.

Combining AI with molecular property prediction has the potential to accelerate the discovery of new materials [80], the development of sustainable materials [81], and the design of targeted drugs [82]. This integration will effectively tackle challenges faced by the chemical industry while achieving goals related to economy, efficiency, and sustainability [83]. It is anticipated that a deeper integration of AI with chemical industry knowledge and technology will create new opportunities for scientific research [60] and industrial innovation [84], leading to more accurate and diverse applications [85].

3 Synergistic optimization of molecular generation and process design

The process of selecting and designing suitable molecules for special functions is a typically time-consuming and labor-intensive task [86]. There are two main approaches for molecule design: experimental trial-and-error methods [87] and computer-aided molecular design (CAMD) [88]. Compared to CAMD, experimental trial-and-error approaches are resource-intensive, pose safety hazards, and fall short of meeting modern industrial and environmental requirements. Hence, it is of critical importance to shift from traditional empirical trial-and-error experiments to experiments guided by theoretical prediction in order to achieve the desired goal with minimal resources [89].

CAMD is a comprehensive term [90] that encompasses a variety of computational methods aimed at the rational design of molecules with specific and optimal properties derived from designated molecular building blocks. It primarily generates new molecules primarily by combining groups [91], which can lead to combinatorial explosions. Additionally, it struggles to distinguish some isomers and separate groups from some complex structures. Another challenge is the extensive scope of the chemical space, which is too vast to be comprehensively considered.

Molecular structure generation can be achieved by constructing an encoder and a decoder to form a molecular generative model [92]. However, the model generates a large number of molecules with varying performances. When designing molecules for specific functions, it becomes necessary to implement constrained screening and targeted structure generation steps to obtain the desired molecules that show excellent performance in processes (Fig. 3). The following section will provide a detailed discussion of cases involving constrained screening and targeted structure generation.

Constraint screening ensures that the designed molecules are more readily synthesizable and better aligned with practical needs. As a result, the ultimately selected molecules are more likely to be applicable in practical scenarios. The screening criteria generally focus on performance and process aspects. For example, in the separation of cyclohexane and benzene mixtures, the authors identified 1103 promising molecules that demonstrate excellent selectivity and solvation capacity as green extractive solvents [93]. Subsequently, the study assessed the practicality of molecular synthesis using a synthesis score based on fragment contributions and complexity. It was found that most molecules had favorable scores for easy synthesis. Additionally, the model calculations were performed to generate solvent candidates while considering factors like melting and boiling points, along with flash points to ensure safety and economic viability. This process resulted in five retained candidate molecules: 5-methyl furfural, 3-propoxypropanenitrile, 3-but-3-enoxypropanenitrile, 1-phenylpropan-2-one, and 2-phenyl-1,4-dioxane.

For the purpose of seeking better-performing solvent molecules, residue curve maps were further calculated to assess whether a solvent is suitable for extractive distillation to separate a mixture. This was done using Aspen Plus to calculate the energy consumption for both the extractive and regeneration columns. Importantly, when compared to four widely used industrial extractive solvents, 5-methyl furfural shows no significant increase in energy consumption. Based on these analyses, 5-methyl furfural can be considered one of the most effective green extractive solvents for the separation of

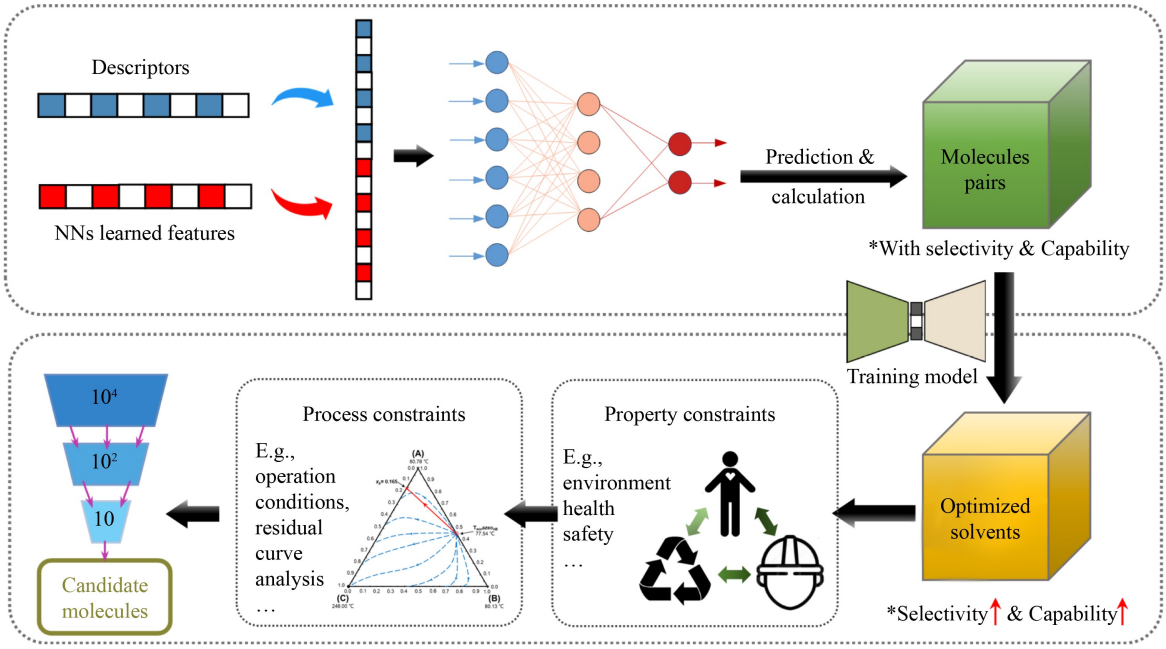


Fig. 3 The conceptual diagram of synergistic optimization of molecular generation and process design.

cyclohexane and benzene. It is worth mentioning that the proposed framework can also be applied to other separation processes, such as liquid-liquid extraction, gas absorption, and crystallization.

Targeted structure generation can be categorized into two types: property-directed and process-required. Zhang et al. [94] also took into account factors such as energy consumption, analysis based on knowledge of the chemistry domain, and molecular fragment analysis. A pre-prepared molecular set that exhibited enhanced selectivity and solvation capacity was employed. A multi-level molecular generation model was developed to explore the optimization pathways for these molecular pairs. The approach enables the training of a multi-objective optimization generation model for molecules. Subsequently, five commonly used solvents in the extraction and distillation separation process of benzene and cyclohexane were optimized, generating 20 optimized solvent molecules for each solvent, resulting in a total of 100 optimized solvent molecules. The optimized solvent molecules were then evaluated based on EH&S properties. In assessing the environmental impact, three key ecological indicators were considered, including factors such as biodegradability, aquatic toxicity, and ozone depletion potential. Health properties of the solvents were evaluated through the rat oral dosage, which served as a critical indicator of toxicity. Safety was assessed by examining the flash point, which indicated the temperature at which a solvent can vaporize to form an ignitable mixture in air. As a result of applying these constraints, 10 solvent molecules remained as viable options. These 10 pre-screened solvent candidates, selected based on their physicochemical properties, were further evaluated against key process-specific requirements. The operational conditions were quantified by the normal melting point and boiling point of the solvents, as a higher boiling point may require excessive energy input for vaporization, thereby reducing the overall process efficiency. After a thorough screening of the operational conditions and an analysis of the results from combined residue curve and univolatility analyses, three solvents emerged as the most suitable candidates: 4-methyl furfural, 5-methyl furfural, and 2-hexanone. These solvents not only meet the specified thermal criteria but also possess desired properties that enhance performance in separation processes.

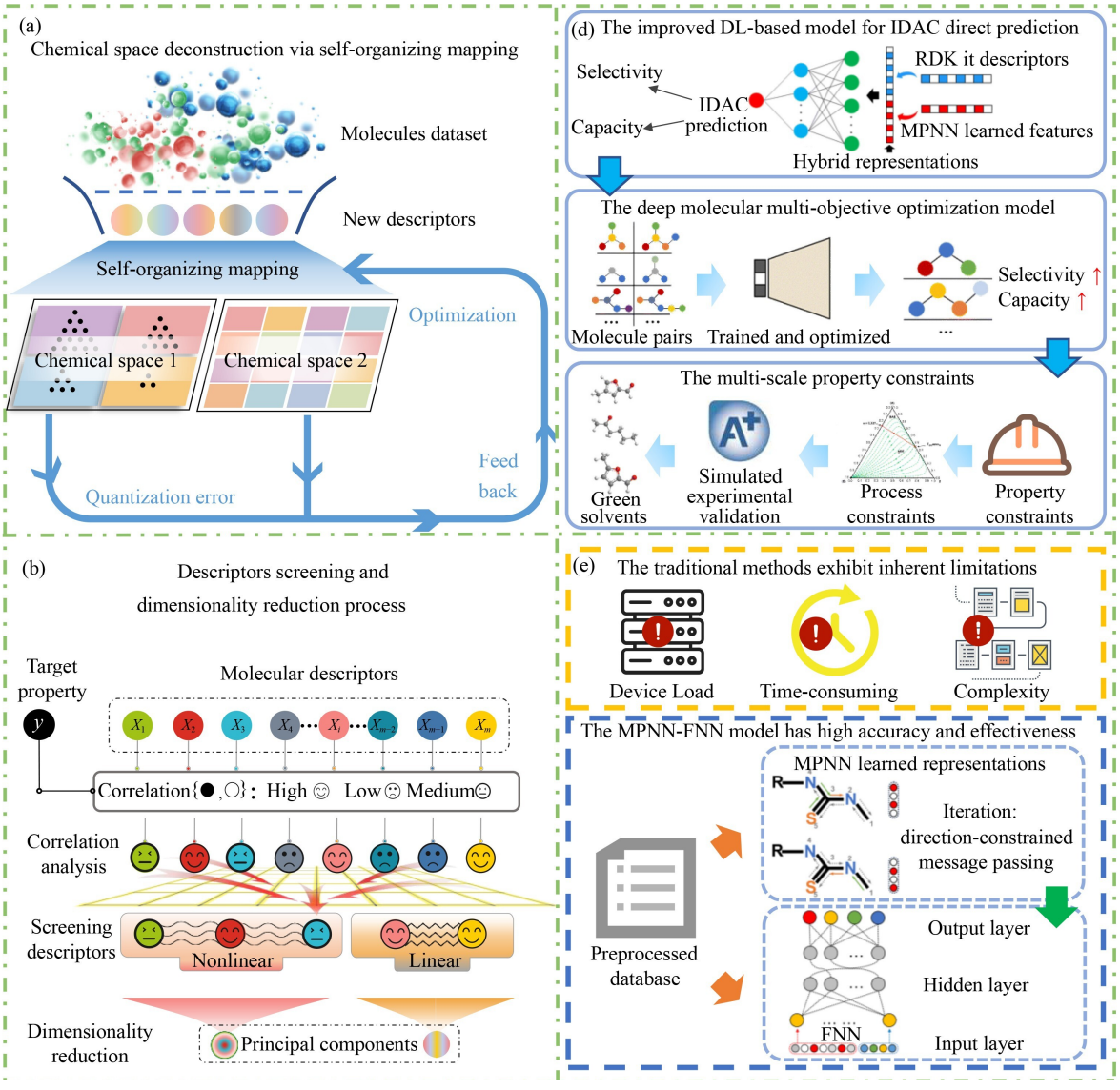
As molecular generation technology advances, molecular design and specific process design will become increasingly interconnected. Advanced computational methods, like DL, make it easier to identify and optimize molecular structures. This allows for the design of substances that are better suited to chemical processes, leading to more innovative process design and optimization (Fig. 4). These improvements will enhance product performance and efficiency. Additionally, by introducing environmentally friendly design conditions, researchers can create new materials that align with the principles of green chemistry. This approach will promote sustainable practices in pharmaceuticals [95], materials [96], and other fields, as such contributing to the sustainable development of the chemical industry.

4 Mechanistic and data-driven integration for process systems

Process is a cornerstone of the manufacturing industry,

significantly influencing aspects such as quality, environment, technology, required skills, and competitive advantage in the market [97]. When production processes or working conditions change, it becomes challenging to respond promptly, which can lead to a decline in production efficiency over time, ultimately affecting overall economic performance [98]. Substantial progress has been made in enhancing green energy efficiency [99] and reducing carbon emissions in chemical process technology [100]. Current efforts are focused on integrating AI into chemical processes, and initial results have been promising [101]. As a result, the concept of intelligent processes has emerged. Herein, the process optimization using AI was applied mainly including neural networks-based hybrid surrogate modeling, model optimization and analysis (Fig. 5). Briefly, the hybrid surrogate model integrates three key elements: a first-principle model based on process mechanisms [102], a data flow section with pseudo-experimental data, and an AI module powered by the most efficient algorithm [103].

The current techniques for process optimization include deterministic algorithms [104], stochastic algorithms [105], heuristic algorithms [106], and others. However, conducting comprehensive and detailed optimization research on large-scale production processes is difficult, especially when it comes to exploring the feasible space of operational variables. The deficiency restricts the ability to meet the real-time optimization and adjustment requirements of complex chemical systems. Furthermore, existing meta-heuristic algorithms have certain limitations when applied to the optimization of high-dimensional objective spaces in intricate systems. Cao et al. [107] proposed a novel framework that integrates DL with an evolutionary algorithm (EA) and utilizes it in a large-scale, complex chemical system: the triple CO_2 feed methanol production process. The hybrid framework achieved rapid optimization of four objective functions and significantly reduced the computational resource



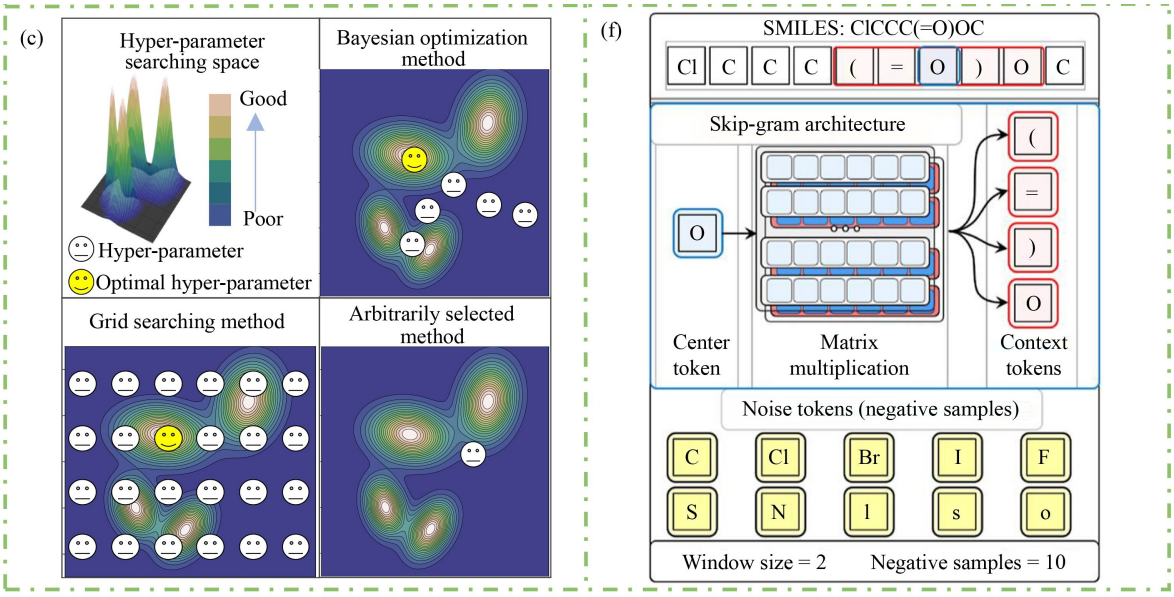


Fig. 4 Recent progress of AI in the chemical industry from the molecular scale. (a) A model ensemble architecture for molecular property prediction. Reprinted with permission from Ref. [68], copyright 2024, Elsevier. (b) A systematic approach to developing robust QSPR models for property prediction. Reprinted with permission from Ref. [67], copyright 2022, Wiley. (c) Hyperparameter optimization for molecular property prediction. Reprinted with permission from Ref. [70], copyright 2022, Chemical Industry Press. (d) Multi-objective optimization strategy for green solvent design. Reprinted with permission from Ref. [94], copyright 2024, Royal Society of Chemistry. (e) MDNN for property prediction. Reprinted with permission from Ref. [61], copyright 2022, Elsevier BV. (f) Integrating NLP for molecular representation in QSPR Modeling. Reprinted with permission from Ref. [46], copyright 2023, American Chemical Society.

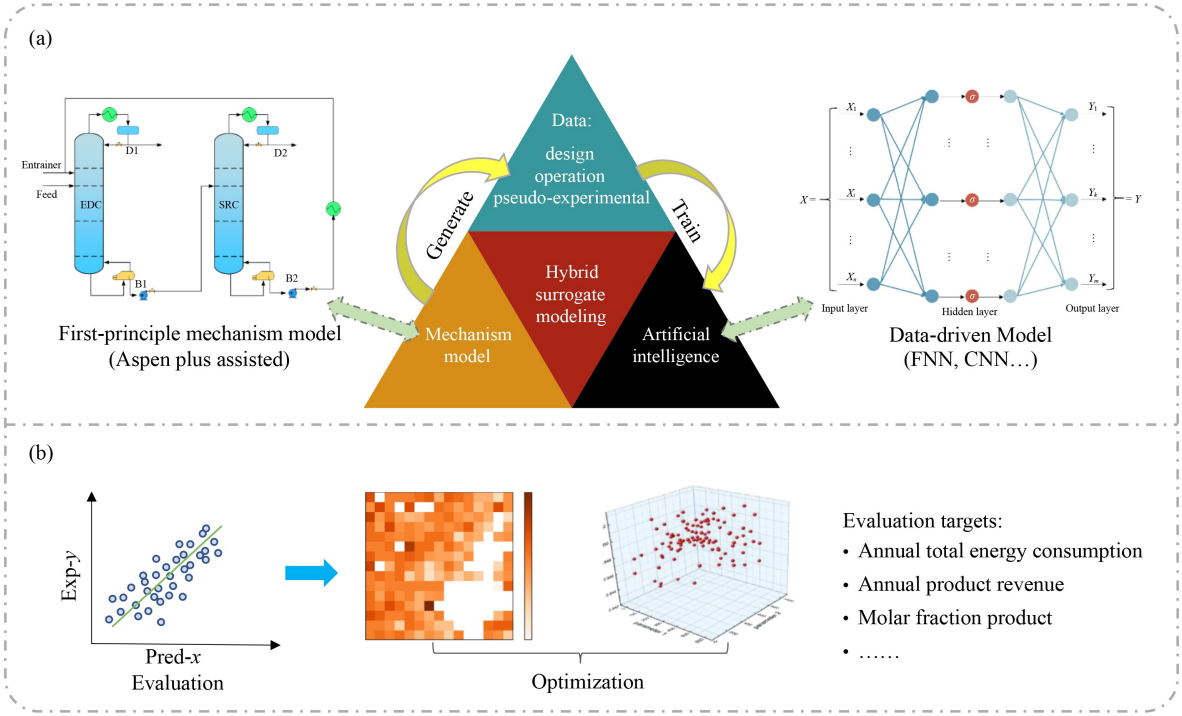


Fig. 5 The conceptual diagram of mechanistic and data-driven integration for process systems. (a) NNs-based hybrid surrogate modeling, (b) model optimization and analysis.

requirements for process optimization, which is helpful to realize real-time optimization of chemical processes.

Chemical process optimization is typically regarded as a multi-objective optimization problem, as it entails balancing various competing objectives, including safety, sustainability, and product quality [108]. Multi-objective optimization is widely applied in chemical engineering [109], particularly in process design, where it facilitates

trade-off analysis and the simultaneous optimization of competing objectives [18]. This enables the identification of Pareto-optimal solutions, supporting efficient resource allocation and enhancing system performance. Multi-objective optimization techniques are primarily classified into two types: evolutionary and swarm-based techniques [110]. EAs draw inspiration from the natural process of biological evolution, whereas swarm-based algorithms are derived from the collective intelligence behaviors observed in biological systems. Compared to EAs, swarm-based techniques generally exhibit faster convergence rates but are more prone to getting trapped in local optima, especially when dealing with high-dimensional or multimodal optimization problems. Consequently, EAs and hybrid approaches that integrate the advantages of both methodologies have attracted significant attention from researchers, driving extensive advancements [111]. In recent years, EAs have witnessed exciting advancements through integrating with cutting-edge technologies and methodologies [112]. One promising direction is the combination of EAs with large-scale ML models, such as deep neural networks and generative models, to enhance optimization efficiency and scalability. Additionally, EAs are increasingly being coupled with interactive decision-making frameworks, where human decision-makers or preference models guide the optimization process. This integration allows for more user-centric solutions, particularly in multi-objective optimization problems, by incorporating real-time feedback and domain-specific knowledge. These developments highlight the adaptability of EAs and their potential to address increasingly complex and high-dimensional challenges in both theoretical and applied contexts.

To solve the complex multi-objective optimization problem of the light olefins separation system, Yang et al. [113] proposed an innovative strategy that integrates artificial neural network-based surrogate models with an enhanced version of the Non-dominated Sorting Genetic Algorithm II. First, a steady-state model of a lightweight olefin separation system was established based on industrial data. Meanwhile, a data set was generated through the MS Excel-Python-Aspen Plus interface [114]. Following this, a process surrogate model was developed using an artificial neural network to evaluate process energy consumption, product yield, and product purity. Finally, the integrating model was used to quickly identify the optimal operating parameters under energy and economic objectives. The optimized operational scheme significantly reduced annual energy consumption while only having a slight impact on annual product revenue. Notably, the algorithm employed the dynamic simulated binary crossover operator and the dynamic polynomial mutation operator to specifically tackle the nonlinear programming problem and the global optimal problem.

The intelligent optimization model enhances computational efficiency and robustness compared to traditional architectures, effectively guiding parameter adjustments in industrial production processes. It tackles the complexities associated with large-scale chemical process modeling and optimization, overcoming issues such as slow convergence, low efficiency, and poor robustness. The proposed multi-objective optimization framework provides production operators with clear guidance on optimal operating parameters, contributing to the sustainable development and clean production of large-scale processes. Ultimately, the model facilitates energy savings and efficiency improvements throughout the entire process. The efficiency of integrating ML with specific process technologies has also been validated on organic Rankine cycles by Zhou et al. [115] and on the vacuum pressure swing adsorption process for capturing CO_2 in enclosed spaces by Du et al. [116].

Multiscale multiphase reactor with AI is a research hotspot that has sparked increasing scientific interest [117]. Reactor is essential to chemical industry, and their intensification, which encompasses aspects such as geometrical shape, configuration, operational conditions, and transport properties, has a significant impact on the overall sustainability of chemical processes [118]. The addition of AI helps researchers gain a better understanding of the complex physical-chemical phenomena in multi-scale (e.g., micro-scale, meso-scale, and macro-scale) and multi-phase (e.g., gas, liquid, and solid) reactors. Developing a hybrid data-driven model based on first principles can considerably reduce the time and cost involved in the design of optimal reactors. Lei et al. [119] innovatively integrated a computational fluid dynamics model with ML to develop a sub-grid filter model for closed gas-solid momentum transfer, which is applicable to industrial reactor simulation. The research focuses on elucidating the non-uniform structural characteristics of reactors and uncovering the fundamental principles governing reactor design and scale-up processes, thereby offering valuable theoretical insights for reactor development. Subsequent work will emphasize AI-assisted numerical simulations to predict key reactor parameters, along with constructing microscale closure and mesoscale correction models. Looking ahead, AI in multiscale multiphase reactors shows great potential for improving predictive accuracy and operational efficiency, paving the way for more advanced and adaptable reactor designs in the future.

In summary, the combination of AI technology with process simulation systems can optimize chemical processes by understanding the underlying mechanisms. This approach helps in creating more accurate prediction models, which can offer reliable insights for specific projects. Additionally, integrating multi-scale and multi-disciplinary data will foster interdisciplinary collaboration, enabling the chemical industry to respond more

flexibly and efficiently to complex challenges.

5 Dynamic digital twin modeling and optimization at production system scale

In the actual production process, the traditional quality control technology faces several issues, including low traceability efficiency, control lag, and unpredictability. Digital twin technology, on the other hand, can realize the quality control of dynamic processes by establishing a high-fidelity model [120], and it has gradually become a hot topic in the field of intelligent processing [121].

Li et al. [122] integrated the principles of chemical processes with deep neural networks to develop a light-attention-mixed-base DL architecture (LAMBDA) that simultaneously accomplishes process knowledge learning and high-accuracy multivariable modeling. LAMBDA consists of three submodules: the multi-kernel convolution module, the light attention module, and the residual module. The multi-kernel convolution module plays a vital role in preventing overfitting and to identify the dynamic characteristics of the system processes. The light attention module corrected deviations in fundamental dynamic characteristics that were caused by transient disturbances. The residual module was designed to reduce interference affecting the target input. The DL architecture demonstrates remarkable accuracy and robustness. It incorporates multitask learning and interpretable ML techniques. This architecture not only plays a facilitating role in enabling multivariable modeling of processes but also provides valuable assistance in analyzing actual chemical processes. Significantly, it can accommodate arbitrary output quantities without any negative impact on its performance. The framework holds substantial promise in terms of its applications, as it makes important contributions to the development of sophisticated intelligent modeling frameworks. Moreover, it offers valuable insights that are instrumental in the creation and improvement of advanced monitoring and control strategies. With the aim of achieving long-term multi-step forecasting, it will be valuable to conduct more in-depth investigations into the integration of LAMBDA with various DL methodologies.

In the realm of chemical engineering, multi-step prediction of variable trends serves as the core foundation for numerous online applications [123]. Data-driven modeling is widely used as one of the most prevalent tools for predicting variable trends [124]. However, there are still notable deficiencies when it comes to the discovery of correlation analysis techniques that are based on the intrinsic correlation of variables. Up to now, no correlation analysis techniques have been found to be applicable to unequal-length variable sequences. To bridge this research gap, our research group has achieved significant progress in the area of dynamic process forecasting by conducting correlation analysis on sequences of unequal lengths [125]. Specifically, we developed a correlation-similarity coupling algorithm (CSCA) based on chemical process mechanisms. This algorithm is highly effective in identifying the most relevant features while minimizing redundancy. Furthermore, we have also constructed an LSTM encoder-decoder DL multi-step prediction model based on CSCA, which enabled multi-step predictions of various process parameters involved in actual de-ethanization industrial processes (Fig. 6). The framework exhibits significant potential for further development and can be integrated with other AI technologies. Through such integration, enhanced process control outcomes can be achieved, thereby making a significant contribution to online applications within the chemical industry. This integration also promotes the advancement of intelligent chemical engineering. Additional research efforts will focus on introducing more effective noise reduction modules to improve model robustness and validate the new architecture on more chemical processes.

As intelligent manufacturing continues to evolve, we

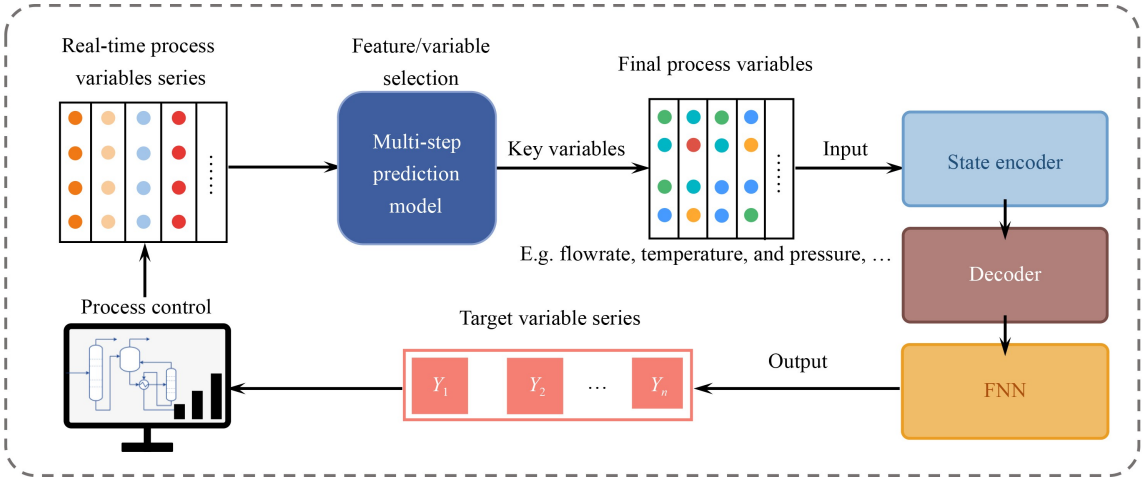
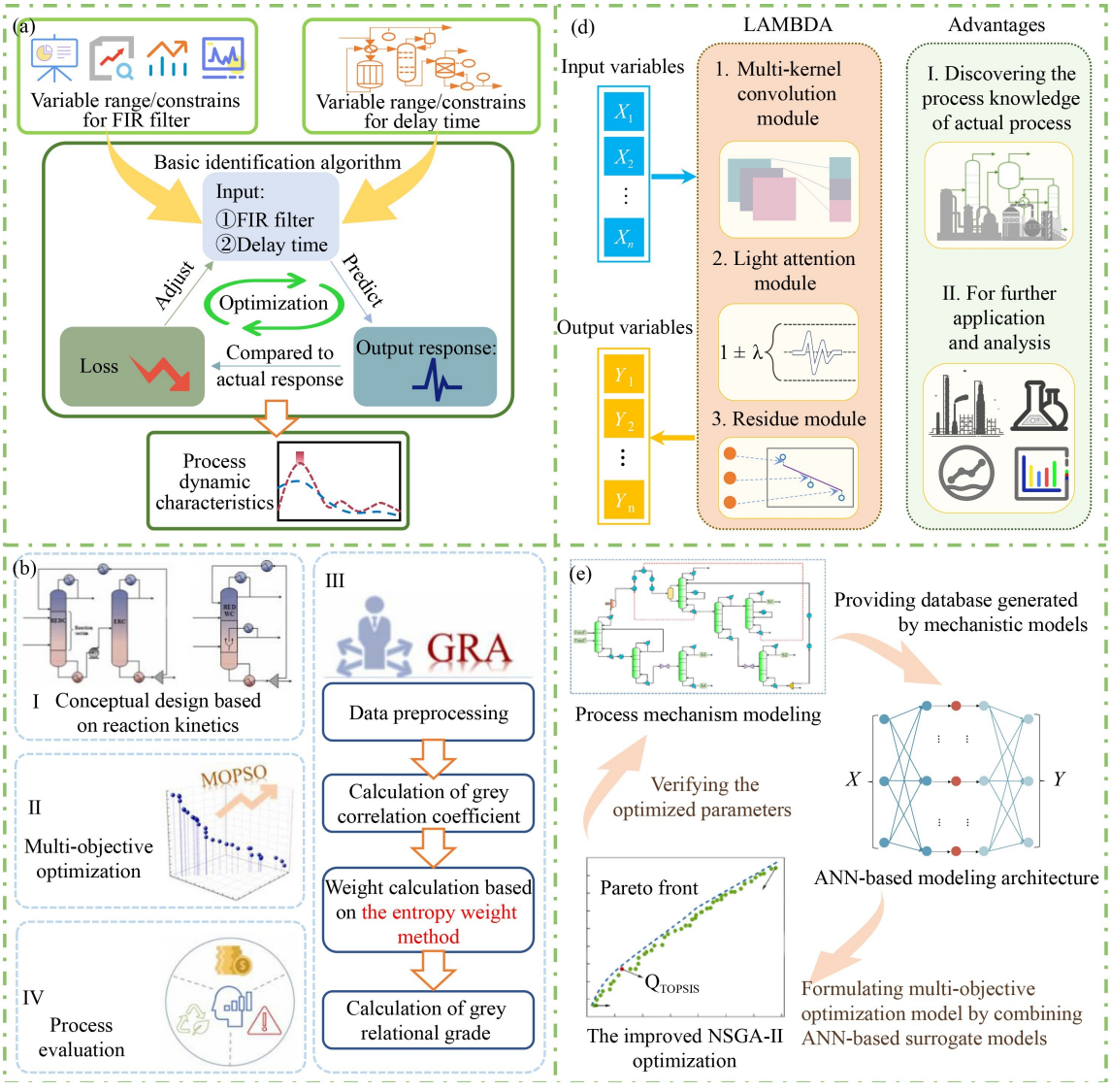


Fig. 6 Real-time optimization of process parameters based on multi-step prediction models.

anticipate that the maturity of dynamic digital twin modeling technology will be achieved through the combination of advanced AI-related technologies. This will, in turn, facilitate the promotion of real-time monitoring and optimization of chemical processes. It will also enable long-term prediction of processes and provide more accurate decision support (Fig. 7). Moreover, we expect that dynamic digital twin modeling technology will assume an increasingly significant role within various intelligent production systems. It will assist chemical industries in achieving digital transformation and realizing intelligent operations.

6 Conclusions and perspectives

As has been elaborated above, data-driven and mechanism-based intelligent techniques are gradually emerging as the new driving forces for addressing the practical challenges encountered in the chemical industry. Meanwhile, the effective utilization and enhancement of these techniques act as catalysts for resolving these challenges. The integration of these methodologies encompasses a wide range of applications, including but not limited to molecular property prediction, high-throughput screening, and multi-objective solvent generation. Moreover, these approaches contribute to significant advancements in both molecular structure optimization and global process optimization. A groundbreaking evolution is observed in the transition from low-dimensional molecular representations to high-dimensional, multimodal feature extraction, as well as the progression from single-objective to multi-objective predictive models. Additionally, there has been a marked transformation from conventional black-box models to more sophisticated intelligent models, which strive to achieve an optimal balance between interpretability and predictive accuracy. This extends to the development of systematic modeling frameworks capable of quantifying uncertainty and precisely defining application boundaries,



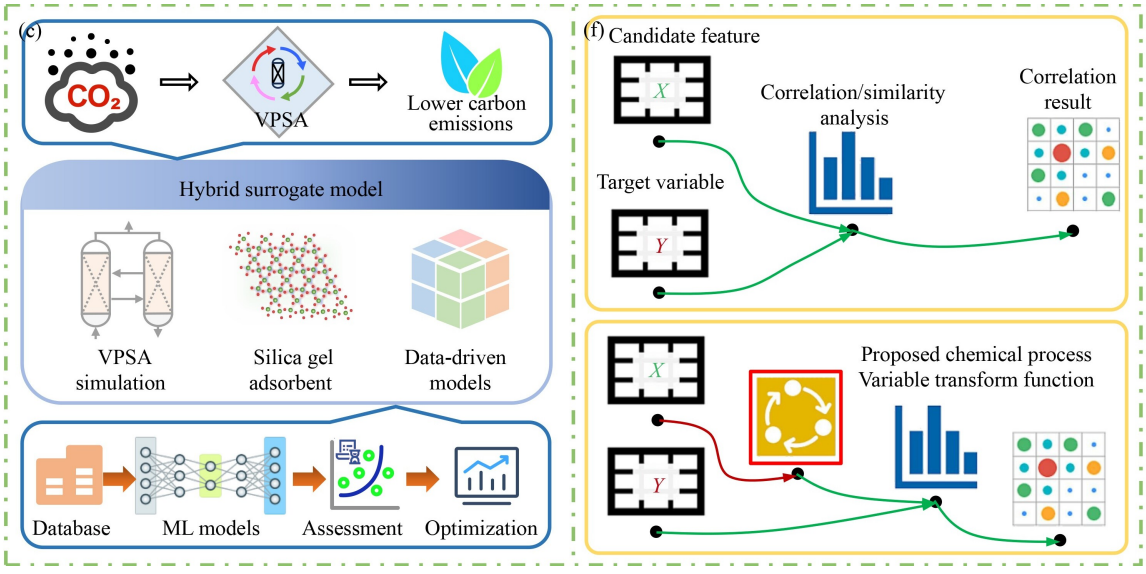


Fig. 7 Recent progress of AI in the chemical industry from the factory scale. (a) A mechanism-data hybrid-driven framework for chemical processes. Reprinted with permission from Ref. [99], copyright 2023, Institution of Chemical Engineers. (b) Multi-criteria optimization for reactive extractive distillation. Reprinted with permission from Ref. [100], copyright 2023, Elsevier. (c) A hybrid surrogate model structure for design and optimization of CO_2 capture processes. Reprinted with permission from Ref. [116]. (d) A light attention-mixed-base DL architecture for process modeling. Reprinted with permission from Ref. [122], copyright 2023, Elsevier. (e) An ML-driven multi-objective optimization architecture for a light olefins separation system. Reprinted with permission from Ref. [113], copyright 2024, Elsevier. (f) A correlation-similarity conjoint algorithm for developing a multi-step prediction model of the chemical process. Reprinted with permission from Ref. [125], copyright 2024, Elsevier.

moving beyond simplistic modeling approaches.

Nevertheless, the performance of models is restricted by several factors, including data quality, complexity, computational resource requirements, interpretability, and a lack of domain knowledge. These limitations curtail their practical applications to a certain extent. In response, researchers are increasingly devoting efforts to optimize models, algorithms, and feature engineering. They are integrating diverse technological methods to augment robustness and interpretability. This is achieved by adding functional modules and incorporating domain-specific knowledge into existing models. Through these endeavors, the aim is to empower molecular design across multiple scales using AI, gradually developing intelligent processes that contribute to the establishment of smart factories.

Despite the remarkable advancements that AI has witnessed in the chemical industry, a number of challenges still persist. Future research will primarily concentrate on the following directions: the AI-driven optimization of chemical processes and intelligent control, AI-assisted new discovery and personalized design of materials, the deep integration of AI and green chemistry, and the application of AI in chemical safety and risk prediction.

Overall, the construction of intelligent chemical engineering systems represents a monumental task that demands multidimensional and multi-scale modeling strategies. These strategies need to cover the entire spectrum from molecules to processes and systems, with the ultimate objective of building a digital twin platform for industry-related issues. Such a digital twin platform will integrate multimodal and multiscale big data with chemical mechanisms and engineering principles. In doing so, it will offer a plethora of new research paradigms for the development of low-carbon processes. From a broader perspective, it will function as a powerful engine for cross-scale and cross-temporal intelligent collaborative optimization, aiming to maximize the economic, environmental, and social benefits within the chemical industries.

Competing interests The authors declare that they have no competing interests.

Acknowledgements We acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 22278044); the Chongqing Science Fund for Distinguished Young Scholars (Grant No. CSTB2022NSCQ-JQX0021); the Fundamental Research Funds for the Central Universities (Grant No. 2024IAIS-QN004); the Chongqing Innovation Support Key Program for Returned Overseas Chinese Scholars (Grant No. CX2023002); the Key Project of Technical Innovation and Application Development (Grant No. CSTB2024TIAD-KPX0058); the Science and Technology Innovation Key R&D Program of Chongqing (Grant No. CSTB2024TIAD-STX0032); the Xinjiang Autonomous Region Regional Collaborative Innovation Special Science and Technology Assistance Plan Project (Grant No. 2024E02036); the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2024B01).

References

1. Daoutidis P, Megan L, Tang W. The future of control of process

systems. Computers & Chemical Engineering, 2023, 178: 108365
2. He C, Zhang C, Bian T, Jiao K, Su W, Wu K J, Su A. A review on artificial intelligence enabled design, synthesis, and process optimization of chemical products for industry 4.0. Processes, 2023, 11(2): 330
3. Wang X, Yu N, Jiao Z, Li L, Yu H, Wei S. Machine learning: enhanced molecular network reveals global exposure to hundreds of unknown PFAS. Science Advances, 2024, 10(21): eadn1039
4. Song J, Han B. Green chemistry: a tool for the sustainable development of the chemical industry. National Science Review, 2015, 2(3): 255–256
5. Qian F. Smart and optimal manufacturing: the key for the transformation and development of the process industry. Engineering, 2017, 3(2): 151
6. Kamkar M, Leonard K C, Ferrer I, Loo S C J, Biddinger E J, Brady D, Carrier D J, Gathergood N, Han H, Hermans I, et al. Artificial intelligence (AI) for sustainable resource management and chemical processes. ACS Sustainable Chemistry & Engineering, 2024, 12(8): 2924–2926
7. Zhou T, Mcbride K, Linke S, Song Z, Sundmacher K. Computer-aided solvent selection and design for efficient chemical processes. Current Opinion in Chemical Engineering, 2020, 27: 35–44
8. Lai N, Tew Y, Zhong X, Yin J, Li J, Yan B, Wang X. Artificial intelligence (AI) workflow for catalyst design and optimization. Industrial & Engineering Chemistry Research, 2023, 62(43): 17835–17848
9. Lee J W, Kim S, Pineda I T. Review of nanoabsorbents for capture enhancement of CO_2 and its industrial applications with design criteria. Renewable & Sustainable Energy Reviews, 2021, 138: 110524
10. Deng W, Liu L, Li X, Huang Y, Hu M, Zheng Y, Yin Y, Huan Y, Cui S, Sun Z, et al. Machine-learning-enhanced trial-and-error for efficient optimization of rubber composites. Advanced Materials, 2025, 37(16): 2407763
11. Chen C, Reniers G. Chemical industry in China: the current status, safety problems, and pathways for future sustainable development. Safety Science, 2020, 128: 104741
12. Zhao J, Shi T, Ren J. Accelerating operation optimization of complex chemical processes: a novel framework integrating artificial neural network and mixed-integer linear programming. Chemical Engineering Journal, 2024, 481: 148421
13. Yang T, Yi X, Lu S, Johansson K H, Chai T. Intelligent manufacturing for the process industry driven by industrial artificial intelligence. Engineering, 2021, 7(9): 1224–1230
14. Tiong L C O, Yoo H J, Kim N, Kim C, Lee K Y, Han S S, Kim D. Machine vision-based detections of transparent chemical vessels toward the safe automation of material synthesis. npj Computational Materials, 2024, 10(1): 42
15. Zhou G, Zhang C, Li Z, Ding K, Wang C. Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing. International Journal of Production Research, 2020, 58(4): 1034–1051
16. de Almeida A F, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. Nature Reviews. Chemistry, 2019, 3(10): 589–604
17. Yang X, Zhou J, Xie Z, Ke G. Chemical process fault diagnosis based on enchanted machine-learning approach. Canadian Journal of Chemical Engineering, 2019, 97(12): 3074–3086
18. Xu W, Wang Y, Zhang D, Yang Z, Yuan Z, Lin Y, Yan H, Zhou X, Yang C. Transparent AI-assisted chemical engineering process: machine learning modeling and multi-objective optimization for integrating process data and molecular-level reaction mechanisms. Journal of Cleaner Production, 2024, 448: 141412
19. Jiang S, Zavala V M. Convolutional neural nets in chemical engineering: foundations, computations, and applications. AIChE Journal, 2021, 67(9): e17282
20. Chiang L H, Braun B, Wang Z, Castillo I. Towards artificial intelligence at scale in the chemical industry. AIChE Journal, 2022, 68(6): e17644
21. Tao F, Xiao B, Qi Q, Cheng J, Ji P. Digital twin modeling. Journal of Manufacturing Systems, 2022, 64: 372–389
22. Li J, Zhao T, Yang Q, Du S, Xu L. A review of quantitative structure-activity relationship: the development and current status of data sets, molecular descriptors, and mathematical models. Chemometrics and Intelligent Laboratory Systems, 2025, 256: 105278
23. Chen M, Yang J, Tang C, Lu X, Wei Z, Liu Y, Yu P, Li H H. Improving ADMET prediction accuracy for candidate drugs: factors to consider in QSPR modeling approaches. Current Topics in Medicinal Chemistry, 2024, 24(3): 222–242
24. Zhu T, Li S, Tao C. A new perspective on predicting the reaction rate constants of hydrated electrons for organic contaminants: exploring molecular structure characterization methods and ambient conditions. Science of the Total Environment, 2023, 904: 166316
25. Ma M, Lei X. A deep learning framework for predicting molecular property based on multi-type features fusion. Computers in Biology and Medicine, 2024, 169: 107911
26. Liang X, Zhang X, Zhang L, Liu L, Du J, Zhu X, Ng K M. Computer-aided polymer design: integrating group contribution and molecular dynamics. Industrial & Engineering Chemistry Research, 2019, 58(34): 15542–15552
27. Shaahmadi F, Smith S A, Schwarz C E, Burger A J, Cripwell J T. Group-contribution SAFT equations of state: a review. Fluid Phase Equilibria, 2023, 565: 113674
28. Hu Y, Su Y, Jin S, Chien I L, Shen W. Systematic approach for screening organic and ionic liquid solvents in homogeneous extractive distillation exemplified by the *tert*-butanol dehydration. Separation and Purification Technology, 2019, 211: 723–737
29. Cocchi M, Menziani M C, Debenedetti P G, Cruciani G. Theoretical versus empirical molecular descriptors in monosubstituted benzenes. Chemometrics and Intelligent Laboratory Systems, 1992, 14(1-3): 209–224
30. Mauri A, Consonni V, Todeschini R. Molecular Descriptors. Berlin: Springer, 2017, 2065–2093
31. Zhang R, Wang Y, Zhu W, Xin L, Qi J, Wang Y, Zhu Z, Cui P. Insight into the mechanism of machine learning models for

predicting ionic liquids toxicity based on molecular structure descriptors. ACS Sustainable Chemistry & Engineering, 2024, 12(49): 17749–17760

32. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chemical Reviews, 2019, 119(18): 10520–10594
33. Zhang K, Yang X, Wang Y, Yu Y, Huang N, Li G, Li X, Wu J C, Yang S. Artificial intelligence in drug development. Nature Medicine, 2025, 31(1): 45–59
34. Dong J, Cao D, Miao H, Liu S, Deng B C, Yun Y H, Wang N N, Lu A P, Zeng W B, Chen A F. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. Journal of Cheminformatics, 2015, 7(1): 60
35. Zhu Z, Wu D, Zhang J, Ren J, Jin S, Shen W. An adaptive and interpretable modeling architecture assisted rapid and reliable consensus prediction for hazardous properties of chemicals. Journal of Cleaner Production, 2024, 471: 143441
36. Wu Z, Jiang D, Wang J, Zhang X, Du H, Pan L, Hsieh C Y, Cao D, Hou T. Knowledge-based BERT: a method to extract molecular features such as computational chemists. Briefings in Bioinformatics, 2022, 23(3): bbac131
37. Xu J, Wang L, Wang L, Liang G, Shen X, Xu W. Prediction of Setschenow constants of organic compounds based on a 3D structure representation. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 178–184
38. Xiang H, Zeng L, Hou L, Li K, Fu Z, Qiu Y, Nussinov R, Hu J, Rosen-Zvi M, Zeng X, et al. A molecular video-derived foundation model for scientific drug discovery. Nature Communications, 2024, 15(1): 9696
39. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nature Machine Intelligence, 2021, 3(12): 1023–1032
40. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design, 2016, 30(8): 595–608
41. Wu D, Zhu Z, Zhang J, Wen H, Jin S, Shen W. An interpretable solute-solvent interactive attention module intensified graph-learning architecture toward enhancing the prediction accuracy of an infinite dilution activity coefficient. Industrial & Engineering Chemistry Research, 2024, 63(19): 8741–8750
42. Zhang J, Wang Q, Lei Y, Shen W. An interpretable 3D multi-hierarchical representation-based deep neural network for environmental, health, and safety properties prediction of organic solvents. Green Chemistry, 2024, 26(7): 4181–4191
43. Hirschberg J, Manning C D. Advances in natural language processing. Science, 2015, 349(6245): 261–266
44. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018, 13(3): 55–75
45. Su Y, Wang Z, Jin S, Shen W, Ren J, Eden M. An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. AIChE Journal, 2019, 65(9): e16678
46. Zhou Z, Eden M, Shen W. Treat molecular linear notations as sentences: accurate quantitative structure-property relationship modeling via a natural language processing approach. Industrial & Engineering Chemistry Research, 2023, 62(12): 5336–5346
47. Li J, Cheng K, Wang S, Morstatter F, Trevino R P, Tang J, Liu H. Feature selection: a data perspective. ACM Computing Surveys, 2018, 50(6): 1–45
48. Ponzoni I, Sebastián-Pérez V, Requena-Triguero C, Roca C, Martínez M J, Cravero F, Díaz M F, Páez J A, Arrayás R G, Adrio J, et al. Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery. Scientific Reports, 2017, 7(1): 2403
49. Yang A, Sun S, Su Y, Kong Z, Ren J, Shen W. Insight to the prediction of CO_2 solubility in ionic liquids based on the interpretable machine learning model. Chemical Engineering Science, 2024, 297: 120266
50. Wen H, Nan S, Wu D, Sun Q, Tong Y, Zhang J, Jin S, Shen W. A systematic review on intensifications of artificial intelligence assisted green solvent development. Industrial & Engineering Chemistry Research, 2023, 62(48): 20473–20491
51. Huang B, Tong Y, Chen Y, Eslamimanesh A, Shen W, Wei S. Dual self-adaptive intelligent optimization of feature and hyperparameter determination in constructing a DNN based QSPR property prediction model. Industrial & Engineering Chemistry Research, 2022, 61(32): 12052–12060
52. Wang Z, Su Y, Jin S, Shen W, Ren J, Zhang X, Clark J. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. Green Chemistry, 2020, 22(12): 3867–3876
53. Kasza J, Wolfe R. Interpretation of commonly used statistical regression models. Respirology, 2014, 19(1): 14–21
54. Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nature Reviews Drug Discovery, 2024, 23(2): 141–155
55. Matsuzaka Y, Uesawa Y. Ensemble learning, deep learning-based, and molecular descriptor-based quantitative structure-activity relationships. Molecules, 2023, 28(5): 2410
56. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges, and trends. Neurocomputing, 2020, 408: 189–215
57. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of Computer-Aided Molecular Design, 2020, 34(10): 1013–1026
58. Ahmad I, Basheri M, Iqbal M J, Rahim A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access: Practical Innovations, Open Solutions, 2018, 6: 33789–33795
59. Song Z, Shi H, Zhang X, Zhou T. Prediction of CO_2 solubility in ionic liquids using machine learning methods. Chemical Engineering Science, 2020, 223: 115752
60. Baum Z J, Yu X, Ayala P Y, Zhao Y, Watkins S P, Zhou Q. Artificial intelligence in chemistry: current trends and future directions. Journal of Chemical Information and Modeling, 2021, 61(7): 3197–3212
61. Zhang J, Wang Q, Shen W. Message-passing neural network

based multi-task deep-learning framework for COSMO-SAC based σ-profile and VCOSMO prediction. Chemical Engineering Science, 2022, 254: 117624

62. Wang Z, Su Y, Shen W, Jin S, Clark J, Ren J, Zhang X. Predictive deep learning models for environmental properties: the direct calculation of octanol-water partition coefficients from molecular graphs. Green Chemistry, 2019, 21(16): 4555–4565
63. 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, 2021, 7(9): 1201–1211
64. Cao Y, Taghvaie Nakhjiri A, Ghadiri M. Different applications of machine learning approaches in materials science and engineering: comprehensive review. Engineering Applications of Artificial Intelligence, 2024, 135: 108783
65. Wang J, Liu K, Zhang Y, Leng B, Lu J H. Recent advances of few-shot learning methods and applications. Science China Technological Sciences, 2023, 66(4): 920–944
66. Jiang W, Huang K, Geng J, Deng X. Multi-scale metric learning for few-shot learning. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(3): 1091–1102
67. Wen H, Su Y, Wang Z, Jin S, Ren J, Shen W, Eden M. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. AIChE Journal, 2022, 68(1): e17402
68. Wen H, Nan S, Zhang J, Lei Z, Shen W. Chemical space deconstruction-based dynamic model ensemble architecture for molecular property prediction. Chemical Engineering Science, 2024, 295: 120118
69. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing, 2020, 415: 295–316
70. Zhang J, Wang Q, Shen W. Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library. Chinese Journal of Chemical Engineering, 2022, 52: 115–125
71. Acar E, Bayrak G, Jung Y, Lee I, Ramu P, Ravichandran S S. Modeling, analysis, and optimization under uncertainties: a review. Structural and Multidisciplinary Optimization, 2021, 64(5): 2909–2945
72. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing & Applications, 2020, 32(24): 18069–18083
73. Zhang Y, Tino P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 5(5): 726–742
74. Carter A, Imtiaz S, Naterer G F. Review of interpretable machine learning for process industries. Process Safety and Environmental Protection, 2023, 170: 647–659
75. Zhang J, Wang Q, Su Y, Jin S, Ren J, Eden M, Shen W. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChE Journal, 2022, 68(6): e17634
76. Yang A, Sun S, Qi L, Kong Z, Sunarso J, Shen W. Development of an interpretable QSPR model to predict the octanol-water partition coefficient based on three artificial intelligence algorithms. Green Chemical Engineering, 2025, 6(2): 193–199
77. Koscher B A, Canty R B, McDonald M A, Greenman K P, McGill C J, Bilodeau C L, Jin W, Wu H, Vermeire F H, Jin B, et al. Autonomous, multiproperty-driven molecular discovery: from predictions to measurements and back. Science, 2023, 382(6677): eadi1407
78. Vithayathil Varghese N, Mahmoud Q H. A survey of multi-task deep reinforcement learning. Electronics, 2020, 9(9): 1363
79. Yang A, Su Y, Wang Z, Jin S, Ren J, Zhang X, Shen W, Clark J. A multi-task deep learning neural network for predicting flammability-related properties from molecular structures. Green Chemistry, 2021, 23(12): 4451–4465
80. Gao C, Min X, Fang M, Tao T, Zheng X, Liu Y, Wu X, Huang Z. Innovative materials science via machine learning. Advanced Functional Materials, 2022, 32(1): 2108044
81. Goswami L, Deka M K, Roy M. Artificial intelligence in material engineering: a review on applications of artificial intelligence in material engineering. Advanced Engineering Materials, 2023, 25(13): 2300104
82. Mullowney M W, Duncan K R, Elsayed S S, Garg N, van der Hooft J J J, Martin N I, Meijer D, Terlouw B R, Biermann F, Blin K, et al. Artificial intelligence for natural product drug discovery. Nature Reviews Drug Discovery, 2023, 22(11): 895–916
83. Liao M, Lan K, Yao Y. Sustainability implications of artificial intelligence in the chemical industry: a conceptual framework. Journal of Industrial Ecology, 2022, 26(1): 164–182
84. Ureel Y, Dobbelaere M R, Ouyang Y, De Ras K, Sabbe M K, Marin G B, Van Geem K M. Active machine learning for chemical engineers: a bright future lies ahead! Engineering, 2023, 27: 23–30
85. Yu B, Zhang L, Ye X, Wu J Q, Ying H Y, Zhu W, Yu Z Y, Wu X M. State-of-the-art review on various applications of machine learning techniques in materials science and engineering. Chemical Engineering Science, 2025, 306: 121147
86. Austin N D, Sahinidis N V, Trahan D W. Computer-aided molecular design: an introduction and review of tools, applications, and solution techniques. Chemical Engineering Research & Design, 2016, 116: 2–26
87. Millini R. Beyond trial and error for zeolite catalysts. Science, 2017, 355(6329): 1028
88. Martínez T J. *Ab initio* reactive computer aided molecular design. Accounts of Chemical Research, 2017, 50(3): 652–656
89. Zhang Y, Wei S, Dong C, Wang B, Liang Y, Wang Y, Chen X. Progress in QSPR modelling methods. Chinese Science Bulletin, 2021, 66(22): 2832–2844 (in Chinese)
90. Xu S, Cui X, Xing Y, Di J, Zhang X, He J, Feng H. Computer-aided molecular design of double-salt ionic liquid solvents for extractive distillation with the COSMO-SAC and genetic algorithm. Industrial & Engineering Chemistry Research, 2021, 60(49): 18086–18093
91. Chemmangattuvalappil N G. Development of solvent design methodologies using computer-aided molecular design tools. Current Opinion in Chemical Engineering, 2020, 27: 51–59
92. Walters W P, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction.

Accounts of Chemical Research, 2021, 54(2): 263–270
93. Zhang J, Wang Q, Eden M, Shen W. A deep learning-based framework towards inverse green solvent design for extractive distillation with multi-index constraints. Computers & Chemical Engineering, 2023, 177: 108335
94. Zhang J, Wang Q, Wen H, Gerbaud V, Jin S, Shen W. Multi-objective optimization strategy for green solvent design via a deep generative model learned from pre-set molecule pairs. Green Chemistry, 2024, 26(1): 412–427
95. Gao X, Ciura K, Ma Y, Mikolajczyk A, Jagiello K, Wan Y, Gao Y, Zheng J, Zhong S, Puzyn T, et al. Toward the integration of machine learning and molecular modeling for designing drug delivery nanocarriers. Advanced Materials, 2024, 36(45): 2407793
96. Wei J, Chu X, Sun X, Xu K, Deng H X, Chen J, Wei Z, Lei M. Machine learning in materials science. InfoMat, 2019, 1(3): 338–358
97. Shen T, Teng L, Hu Y, Shen W. Systematic screening procedure and innovative energy-saving design for ionic liquid-based extractive distillation process. Frontiers of Chemical Science and Engineering, 2023, 17(1): 34–45
98. Yang A, Su Y, Shi T, Ren J, Shen W, Zhou T. Energy-efficient recovery of tetrahydrofuran and ethyl acetate by triple-column extractive distillation: entrainer design and process optimization. Frontiers of Chemical Science and Engineering, 2022, 16(2): 303–315
99. Li Y, Yang Z, Deng X, Li N, Li S, Lei Z, Eslamimanesh A, Jin S, Shen W. A mechanism-data hybrid-driven framework for identifying dynamic characteristic of actual chemical processes. Chemical Engineering Research & Design, 2023, 199: 115–129
100. Du L, Jin S, Yang Z, Sun S, Yang A, Shen W. An efficient multi-criteria decision making for assessing the optimization of reactive extractive distillation in terms of economy, environment, and safety. Chemical Engineering Research & Design, 2023, 197: 838–850
101. Su Y, Lü L, Shen W, Wei S. An efficient technique for improving methanol yield using dual CO_2 feeds and dry methane reforming. Frontiers of Chemical Science and Engineering, 2020, 14(4): 614–628
102. Shah P, Pahari S, Bhavsar R, Kwon J S I. Hybrid modeling of first-principles and machine learning: a step-by-step tutorial review for practical implementation. Computers & Chemical Engineering, 2025, 194: 108926
103. Mora-Mariano D, Flores-Tlacuahuac A. A machine learning approach for the surrogate modeling of uncertain distributed process engineering models. Chemical Engineering Research & Design, 2022, 186: 433–450
104. Domínguez-Barbero D, García-González J, Sanz-Bobi M Á. Twin-delayed deep deterministic policy gradient algorithm for the energy management of microgrids. Engineering Applications of Artificial Intelligence, 2023, 125: 106693
105. Bishnu S K, Alnouri S Y, Al Mohannadi D M. Stochastic algorithm-based optimization using artificial intelligence/ machine learning models for sorption enhanced steam methane reformer reactor. Computers & Chemical Engineering, 2025, 196: 109060
106. Alatas B. ACROA: artificial chemical reaction optimization algorithm for global optimization. Expert Systems with Applications, 2011, 38(10): 13170–13180
107. Cao H, Li Y, Chang C, Zhang X, Yang A, Shen W. A deep learning hybrid framework combining an efficient evolutionary algorithm for complex many-objective optimization of sustainable triple CO_2 feed methanol production. ACS Sustainable Chemistry & Engineering, 2024, 12(17): 6682–6696
108. Torres A I, Ferreira J, Pedemonte M. Machine learning and process systems engineering for sustainable chemical processes: a short review. Current Opinion in Green and Sustainable Chemistry, 2025, 51: 100982
109. Rangaiah G P, Feng Z M, Hoadley A F. Multi-objective optimization applications in chemical process engineering: tutorial and review. Processes, 2020, 8(5): 508
110. Sharma S, Kumar V. A comprehensive review on multi-objective optimization techniques: past, present, and future. Archives of Computational Methods in Engineering, 2022, 29(7): 5605–5633
111. Sharma N, Liu Y. A hybrid science-guided machine learning approach for modeling chemical processes: a review. AIChE Journal, 2022, 68(5): e17609
112. Slim B, Rituparna D, Abhishek G. Recent advances in evolutionary multi-objective optimization. Berlin: Springer, 2017
113. Yang L, Liu S, Chang C, Yang S, Shen W. An efficient and invertible machine learning-driven multi-objective optimization architecture for light olefins separation system. Chemical Engineering Science, 2024, 285: 119553
114. Al-Malah K. Aspen Plus: Chemical Engineering Applications. New Jersey: John Wiley & Sons, 2022
115. Zhou J, Chu Y, Ren J, Shen W, He C. Integrating machine learning and mathematical programming for efficient optimization of operating conditions in organic Rankine cycle (ORC) based combined systems. Energy, 2023, 281: 128218
116. Du J, Cao H, Li Y, Yang Z, Eslamimanesh A, Fakhroleslam M, Mansouri S S, Shen W. Development of hybrid surrogate model structures for design and optimization of CO_2 capture processes: Part I. vacuum pressure swing adsorption in a confined space. Chemical Engineering Science, 2024, 283: 119379
117. Bracconi M. Intensification of catalytic reactors: a synergic effort of multiscale modeling, machine learning, and additive manufacturing. Chemical Engineering and Processing, 2022, 181: 109148
118. Zhu L, Chen X, Ouyang B, Yan W, Lei H, Chen Z, Luo Z. Review of machine learning for hydrodynamics, transport, and reactions in multiphase flows and reactors. Industrial & Engineering Chemistry Research, 2022, 61(28): 9901–9949
119. Lei H, Zhu L, Luo Z. Study of filtered interphase heat transfer using highly resolved CFD-DEM simulations. AIChE Journal, 2021, 67(4): e17121
120. Liu J, Cao X, Zhou H, Li L, Liu X, Zhao P, Dong J. A digital twin-driven approach towards traceability and dynamic control for processing quality. Advanced Engineering Informatics, 2021, 50: 101395
121. Liu S, Bao J, Zheng P. A review of digital twin-driven

machining: from digitization to intellectualization. Journal of Manufacturing Systems, 2023, 67: 361–378

122. Li Y, Hu L, Li N, Shen W. A light attention-mixed-base deep learning architecture toward process multivariable modeling and knowledge discovery. Computers & Chemical Engineering, 2023, 174: 108259
123. Bai Y, Zhao J. A novel transformer-based multi-variable multi-step prediction method for chemical process fault prognosis. Process Safety and Environmental Protection, 2023, 169: 937–947
124. Bishnu S K, Alnouri S Y, Al-Mohannadi D M. Computational applications using data driven modeling in process systems: a review. Digital Chemical Engineering, 2023, 8: 100111
125. Li Y, Cao H, Wang X, Yang Z, Li N, Shen W. A new correlation-similarity conjoint algorithm for developing encoder-decoder based deep learning multi-step prediction model of chemical process. Chemical Engineering Science, 2024, 288: 119748