



Invited review

Chemical agents designed for oilfield development: A new paradigm empowered by artificial intelligence

Keqiang Wei^{1,2}, Minghui Zhou^{3,4}, Jia Huang^{3,4}, Qun Zhang^{1,2,3,4}, Bin Ding^{1,2,3,4}, Weifeng Lyu^{1,2,3,4}^{*}, Mingguo Peng⁵^{*}

¹University of Chinese Academy of Sciences, Beijing 100049, P. R. China

²Institute of Porous Flow & Fluid Mechanics, Chinese Academy of Sciences, Langfang 065007, P. R. China

³State Key Laboratory of Enhanced Oil and Gas Recovery, Beijing 100083, P. R. China

⁴Research Institute of Petroleum Exploration & Development, PetroChina, Beijing 100083, P. R. China

⁵School of Petroleum and Natural Gas Engineering, Changzhou University, Changzhou 213164, P. R. China

Keywords:

Petrochemical
enhanced oil recovery
chemical flooding
chemical agent design
artificial intelligence

Cited as:

Wei, K., Zhou, M., Huang, J., Zhang, Q., Ding, B., Lyu, W., Peng, M. Chemical agents designed for oilfield development: A new paradigm empowered by artificial intelligence. *Advances in Geo-Energy Research*, 2025, 17(1): 1-16.

<https://doi.org/10.46690/ager.2025.07.01>

Abstract:

With the ongoing rise in global energy demand, the importance of enhanced oil recovery in oilfield development is becoming increasingly prominent. However, traditional chemical flooding agents face bottlenecks such as poor adaptability to application environments, unclear coupling mechanisms regarding multiple factors, as well as long research and development cycles. This paper systematically discusses the innovative paradigm of oilfield chemical agent development driven by artificial intelligence and proposes four core technological breakthroughs. Firstly, artificial intelligence-empowered molecular simulation technology can reveal the behavior mechanisms of flooding agents under extreme conditions. Secondly, intelligent molecular design using generative adversarial networks and reinforcement learning breaks through the traditional trial-and-error model. Thirdly, the construction of a data-mechanism dual-driven multi-objective optimization model achieves the collaborative prediction of physicochemical properties, economic benefits and environmental friendliness. Lastly, an integrated system of robotic chemist and high-throughput experimental platforms forms a closed-loop system of “artificial intelligence design - automated synthesis - online detection”, yielding a complete ecosystem. The analysis of the current technological development challenges and future development directions reveals that the artificial intelligence-empowered intelligent Research and Development system is expected to promote the transformation of chemical flooding technology toward efficiency, environmental protection and sustainable development, providing a new standard for intelligent oil and gas field development.

1. Introduction

Crude oil continues to play a central role in the global energy landscape, serving as a primary source of fuel and chemical feedstock for transportation, manufacturing, and daily life (IEA, 2024). As conventional oil reserves decline, reducing production costs and maximizing recovery from

mature fields has become essential for ensuring energy security and sustainable development. Primary and secondary recovery methods typically extract only about one-third of the original oil in place, leaving significant amounts of residual oil trapped in heterogeneous, low-permeability formations (Zerpa et al., 2005; Wang et al., 2023; Yuan et al., 2024). To address this challenge, Enhanced Oil Recovery (EOR)

technologies have gained increasing attention. Among them, chemical flooding is one of the most widely applied and effective techniques (Karimov and Toktarbay, 2023). By injecting chemical agents (such as surfactants, polymers, alkalis, and nanoparticles) chemical flooding aims to reduce oil-water Interfacial Tension (IFT), alter reservoir wettability, and increase sweep efficiency, ultimately mobilizing residual oil under harsh reservoir conditions (Druetta et al., 2019; Lv et al., 2023).

However, the success of chemical flooding is contingent on the stability, efficiency, and environmental compatibility of the injected agents under high-temperature, high-salinity, and high-pressure environments. Polyacrylamide (PAM) and Petroleum Sulfonate (PS) have long been the cornerstone chemicals for flooding applications. While PAM provides viscosity for mobility control, it degrades under extreme reservoir conditions, reducing its effectiveness. Various strategies have been developed to improve PAM's thermal and salt resistance, including incorporation of hydrophobic or rigid monomers (Taylor and Nasr-El-Din, 1998; Sabhapondit et al., 2003; Shi et al., 2022). Nonetheless, the optimization process often involves labor-intensive experimentation, with each modification requiring years to scale from lab to field implementation.

Surfactants based on PS are effective at lowering IFT but suffer from formulation complexity, inconsistent active component content, and susceptibility to formation adsorption. Composite surfactant systems have been explored to mitigate these issues, yet reliable performance still depends on extensive core flooding tests and field trials, prolonging development cycles and increasing costs (Rosen, 2012; Kamal et al., 2017). Furthermore, early formulations failed to account for the potential environmental impact, including the toxicity of residual acrylamide in PAM and the ecological risks posed by PS (Millemann et al., 1982; Xiong et al., 2018). Faced with the increasingly diverse reservoir conditions and stringent environmental requirements, the performance of chemical agents is not only affected by their molecular structure but also by numerous factors such as reservoir environment. Essentially, the design of effective chemical flooding agents is a high-dimensional problem involving the classification, processing and learning of a large amount of data while considering all relevant factors simultaneously. However, the traditional chemical agent development model is often limited to the cycle of "single-performance optimization - experimental correction", lacking a systemic design approach that can quickly adapt to different working environments. For this model, it is difficult to integrate and optimize massive data information and solve high-dimensional problems, thus it fails to effectively balance time and cost.

In contrast, recent advances in artificial intelligence (AI) offer promising solutions. The success of AI applications in fields like the Materials Genome Initiative (Wang et al., 2024) and drug development (Vamathevan et al., 2019; Yang et al., 2019) demonstrates its potential for accelerating innovation in EOR chemical design. Machine Learning (ML) and Deep Learning (DL) models can extract complex structure - property relationships from large datasets. Tools such as Graph Neural Network (GNN), molecular fingerprints, and Variational Au-

toencoder (VAE) enable high-throughput screening and generative molecular design. Moreover, Machine Learning Potential (MLP) and enhanced sampling algorithms now allow for accurate, cross-scale molecular simulations under reservoir conditions, offering insights into kinetic behaviors and interfacial interactions critical to chemical efficacy. Complementing these approaches, automated high-throughput synthesis platforms and robotic chemists enable rapid experimental validation of AI-generated molecular candidates. The integration of AI, molecular simulation, and intelligent experimentation creates a closed-loop framework of "design - synthesis - testing - analysis", streamlining the development cycle of chemical flooding agents. This paradigm has the potential to substantially reduce R&D costs, enhance formulation precision, and enable faster adaptation to complex reservoir environments.

Despite the significant progress in AI-driven material discovery and the optimization in fields such as drug development and materials science, the application of AI in oilfield chemical agent design remains underexplored. While AI technologies like Random Forest (RF), Support Vector Machine (SVM) and Convolutional Neural Networks (CNNs) have shown great promise in other domains, their integration into oilfield chemical development is still in its infancy, which highlights the need for a systematic framework that leverages AI to address the high-dimensional challenges of chemical agent design, optimize formulations and accelerate experimental validation.

The remainder of this paper is structured as follows: Section 2 discusses the AI-driven paradigm for oilfield chemical agent development, focusing on four core technological breakthroughs, including molecular simulation, intelligent molecular design, formulation optimization, and intelligent experimental systems. Section 3 analyzes the current technical challenges and future development directions in this field. Finally, Section 4 provides a conclusion and outlook on the transformative impact of AI on oilfield chemical agent development and its potential to drive sustainable and intelligent oil and gas field operations.

2. The AI-driven paradigm for oilfield chemical agent development

The development of oilfield chemicals (such as surfactants, polymers, and nano-flooding agents) faces significant challenges in harsh reservoir environments marked by high temperature and salinity. Traditional approaches to developing these chemicals are often inefficient, costly, and time-consuming. They also struggle with the "curse of dimensionality", where the complex interplay of performance requirements, molecular structures, and environmental factors makes molecular design extremely difficult. However, recent advancements in high-performance computing and ML technologies have transformed this landscape. These innovations include multi-scale molecular simulation, generative molecular design, data-driven formulation optimization, and intelligent experimental platforms. Collectively, they enable researchers to identify effective oil-displacing and auxiliary chemical agents more efficiently, even within limited timeframes and resource constraints. The following sections provide an in-depth explora-

Table 1. Common AI technologies.

Technology	Description	Application scenarios	Advantages
MLP	DNN-based potential energy surfaces with QM fidelity	Simulating molecular behavior under extreme reservoir conditions	Quantum accuracy at classical cost
Generative models (GAN/VAE/Diffusion)	Learn data distribution to create novel molecular graphs	De-novo flooding agents with target IFT, CMC	Rapidly explores huge chemical space
GNN	Message-passing over molecular graphs	Predict IFT, CMC, adsorption directly from topology	Captures local & global structural features
Tree / Kernel Ensembles (RF, GBRT, SVM)	Decision-tree or kernel ensembles for regression	Quick screening of IFT, contact angle, formulation tuning	High accuracy on small, mixed datasets
RL	Reward-guided molecular refinement loop	Iteratively optimise generated molecules for multi-objective EOR metrics	Self-improves toward desired properties

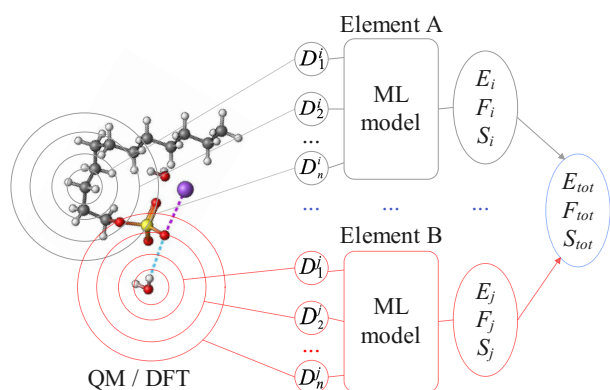


Fig. 1. The basic principle of MLPs. Reproduced with permission from (Kang et al., 2020). Copyright ©2020 American Chemical Society.

tion of four key technological approaches (Sections 2.1-2.4) and examine how they interconnect in practical applications.

Before delving into the specific applications of AI in oilfield chemical agent development, it is essential to briefly introduce some common AI technologies that form the foundation of this paradigm, which are shown in Table 1.

2.1 Molecular simulation

Molecular simulation, such as Molecular Dynamics (MD) and Dissipative Particle Dynamics (DPD), is a key step in the R&D process of oilfield chemicals, which helps to clarify the working mechanisms of additives such as polymers and surfactants in various application environments (Kirch et al., 2020; Santo and Neimark, 2021), conduct research on the structure-property relationship of molecules, and propose directions for molecular design based on the perceived mechanisms. However, traditional simulations face challenges when dealing with complex environments, such as high temperature and salinity, where chemical agents may undergo chemical failure and the chemical reactions involved cannot be solved by traditional molecular simulations. Moreover, despite significant progress in the simulation methods and computing power, existing sim-

ulations continue to face the problem of scale disaster in terms of time and space when dealing with complex working conditions. In other words, quantitative research remains a major challenge at present. To address this, researchers have begun to explore the integration of AI algorithms into the molecular simulation process, forming new technologies such as MLPs (Zhang et al., 2018b; Wen et al., 2022), enhanced sampling strategies, and multi-scale coupling modeling, providing more efficient and accurate ways for mechanism research.

2.1.1 Machine learning potentials

Computational chemistry is an indispensable tool for understanding molecules and predicting their chemical properties. However, due to the difficulties in solving the Schrödinger equation and the increasing computational cost with the size of the molecular system, traditional computational methods face significant challenges (Aldossary et al., 2024).

MLP functions serve as a revolutionary technology bridging the precision of quantum mechanics and the efficiency of classical MD. However, challenges remain in balancing computational cost and accuracy for large-scale systems. As shown in Fig. 1, by training DNNs with high-precision quantum mechanical data, such as Density Functional Theory (DFT) or ab Initio Molecular Dynamics, the model can quickly predict atomic energy and forces, thereby retaining quantum-mechanical precision in large-scale molecular simulations (Shang et al., 2023; Liu et al., 2024). This successfully resolves the long-standing contradiction between computational precision and efficiency in traditional simulation methods.

As an example, the SchNet model by Schütt et al. (2017)'s team innovatively employs continuous filter convolution layers (cfconv) and radial basis function expansion to process atomic distance information, achieving three-dimensional rotational invariance and overcoming the geometric limitations of traditional convolution. On the QM9 dataset, its ground-state energy prediction reaches 0.31 kcal/mol (state-of-the-art), and on the MD17 dataset, the prediction error of atomic forces is only 0.05 kcal/mol/Å. The ISO17 benchmark test shows that with force-supervised training, the prediction error of unknown

molecular energy can be reduced to 2.40 kcal/mol, preliminarily establishing a new paradigm that integrates physical constraints with ML for quantum computation.

The DeePMD series research (Wang et al., 2018; Zhang et al., 2018a; Zeng et al., 2023) uses DNN to learn atomic potential energy functions from data generated by first principle calculations, such as DFT, in order to construct potential energy surfaces. This approach integrates quantum-mechanical precision into molecular dynamics simulations; while maintaining linear computational complexity, it achieves near-first-principle precision (atomic force error of 7.1-19.1 meV/Å), with an efficiency improvement of five orders of magnitude compared to traditional quantum chemical methods.

The upgraded version, DeePMD-kit, employs a three-tier parallel strategy and has successfully completed nanosecond molecular dynamics simulations involving 100 million copper atoms on supercomputers. It achieves a peak performance of 86PFLOPS and maintains 76% parallel efficiency. This represents a two-order-of-magnitude increase in the scale of quantum-level simulations, offering a larger atomic precision platform for mesoscale system research such as amorphous materials (Lu et al., 2021).

In addition, Batzner's team's NequIP model (Batzner et al., 2022) differs from most scalar-invariant convolutional models by employing E(3)-equivariant convolution to handle the interaction of geometric tensors, thereby representing atomic environments more richly and accurately. In the MD-17 benchmark, its force field prediction error is reduced by 78% compared to sGDML, and it demonstrates high data efficiency: Only 133 training samples are needed to achieve the precision of the initial DeepMD using 133,500 samples in water phase change simulations. In the CCSD(T)-level dataset, the force field error of NequIP (3.1 meV/Å) is reduced by 63% compared to that of DimeNet. It successfully reproduces the non-harmonic vibration spectrum of ice and the ion migration barrier of lithium superionic conductors, indicating that networks based on physical symmetries can break through the dependence on large-scale data.

In the field of oilfield chemicals, MLP technology provides a new direction for the structure-property correlation analysis of molecular systems such as surfactants, scale inhibitors and nano-flooding agents. Once a potential energy function with quantum precision is established, researchers can capture the key behaviors of chemical agent molecules applied in oil reservoir flooding in nanosecond mesoscale simulations, providing effective guidance for the optimization design of chemical agent molecules and significantly reducing the experimental verification costs.

2.1.2 AI-driven enhanced sampling

Although molecular dynamics simulations can reveal fine microscopic kinetic processes, for certain rare events that span high energy barriers or slow transitions (such as protein folding, chemical reaction pathways, phase change nucleation, etc.), they often undersample due to limitations in terms of time step and simulation duration. Enhanced sampling techniques (such as metadynamics) accelerate the exploration of low-probability regions or energy barrier crossing processes

by introducing external bias potentials or improving the kinetic equations in the system (Laio and Parrinello, 2002).

With the emergence of AI, the combination of enhanced sampling and ML provides new ideas for multi-scale simulations of complex systems. For instance, in the field of drug design, Bertazzo et al. (2021) proposed a semi-automated computational framework integrating enhanced sampling, ML and physical path analysis to address issues such as high computational cost, insufficient sampling efficiency, and the lack of dynamic path information in the calculation of Absolute Binding Free Energy. By optimizing the ligand dissociation trajectory with the principal path algorithm to construct Path Collective Variables, and integrating metadynamics to reconstruct the free energy surface of the dissociation path, followed by corrections for solvation free energy and configurational entropy, accurate predictions of standard binding free energy can be made. The results show that this method has a good linear correlation with the experimental values in various protein-ligand systems (Pearson correlation coefficients of 0.84 and 0.78), with an average absolute error of about 1.5-2.2 kcal/mol, highlighting the importance of path generation quality for free energy calculations in asymmetric ligand systems.

Targeting the high-precision simulation needs of heterogeneous catalytic systems, Xu et al. (2021) developed the adaptive ML potential-accelerated metadynamics method. It uses Bayesian inference to dynamically assess the variance of potential energy predictions, triggering first principle calculations in real time to update the training set; through a two-stage sampling strategy and Δ -MLP technology (learning the residual energy term on top of the Density Functional Tight Binding base potential), it could achieve free energy surface errors of less than 0.23 and 0.02 eV in the Pt₁₃-CO cluster and Pt (111)-CO surface systems, respectively. The computational efficiency is improved by a factor of 10 compared to traditional DFT-metadynamics, with only 81 DFT calculations required. Based on the scalable base potential function design, this method is compatible with both periodic and non-periodic boundary conditions, providing an efficient and accurate multi-scale simulation tool for the mechanistic study of heterogeneous catalytic reactions.

In high-salinity, high-temperature environments, surfactant molecules may undergo slow structural rearrangements or engage in multiple adsorption mechanisms. To this end, the combination of ML and enhanced sampling can more quickly find energy minima and barriers, as well as analyze the assembly of surfactant molecules and the modification of oil-water interfaces, providing support for the efficient design of surfactants.

2.1.3 AI-driven multi-scale & cross-scale simulation

In the process of chemical flooding with oilfield chemicals, the multi-level physicochemical interactions from molecules to reservoirs are extremely complex. Traditional single-scale models are unable to capture the synergistic effects between femtosecond molecular vibrations and hour-scale displacement processes, nor can they balance the details of angstrom-level molecules with meter-scale reservoir flow behavior (Horstmeier, 2010; Joshi and Deshmukh, 2021; Shilko et al., 2024;

Wang et al., 2025). The hierarchical levels of multi-scale simulation are shown in Fig 2.

In order to break through the limitations of high-precision simulations in temporal and spatial scales, research is evolving along two technical paths: Deep learning-based coarse-grained modeling and adaptive multi-scale coupling modeling. Deep learning-based coarse-grained modeling integrates All-Atom Molecular Dynamics data with deep neural network potential functions to achieve efficient predictions within Coarse-Grained Molecular Dynamics or DPD frameworks. This approach not only enhances the simulation rate by hundreds to thousands of times compared to the original model but also maintains a good description of key molecular interactions, providing new ideas for analyzing the thermodynamic and kinetic characteristics of complex systems (Majewski et al., 2023; Shinkle et al., 2024).

In the field of coarse-grained modeling based on DL, Majewski et al. (2023) proposed a protein modeling method that integrates neural network potential functions, thus developing thermodynamically consistent models based on all-atom molecular dynamics trajectory data. This model, while retaining atomic-scale features, increases the kinetic rate by three orders of magnitude, successfully reconstructing the natural conformation distribution of various proteins and revealing the metastable transition mechanisms in folding pathways through Markov state models. As its breakthrough, it is the first single model integrating 12 different structural proteins that possesses the transfer learning capability for predicting mutant conformations, achieving a synergistic optimization of kinetic fidelity and universality.

Adaptive multi-scale coupling modeling refers to the use of regional decomposition algorithms, a combination of molecular dynamics and Direct Simulation Monte Carlo in nanoscale confined regions (pores < 10 nm or near oil-water interfaces), and modified Navier-Stokes equations for continuous medium modeling in macroscopic pore regions (> 1 μm) (Karniadakis et al., 2005; Tartakovsky and Panchenko, 2016). For example, in shale gas reservoirs, Wang et al. (2020) combined MD with pore network models to more accurately quantify the combined influence of inorganic and organic pores on permeability. The predicted permeability of the model (96.4 ± 11.2 nD) was highly consistent with the experimental values, successfully quantifying the pore structure-transport coupling effect and correcting the traditional bias in permeability assessment that focuses solely on organic matter.

Multi-scale simulation docking technology not only allows for predicting the macroscopic behavior of existing molecules but also provides more complete and realistic evaluation indicators for generative molecular design. With the large number of sample data generated by high-fidelity simulations, AI models can construct more accurate data models that in turn can be used to screen or generate new molecules, forming a closed loop of “simulation-generation-validation”.

Given its powerful capabilities, AI-driven molecular simulation technology has brought about numerous breakthroughs in relevant fields. On the one hand, it can significantly reduce experimental costs and effectively shorten the R&D cycle; on the other hand, it deeply analyzes the cross-scale behavior of

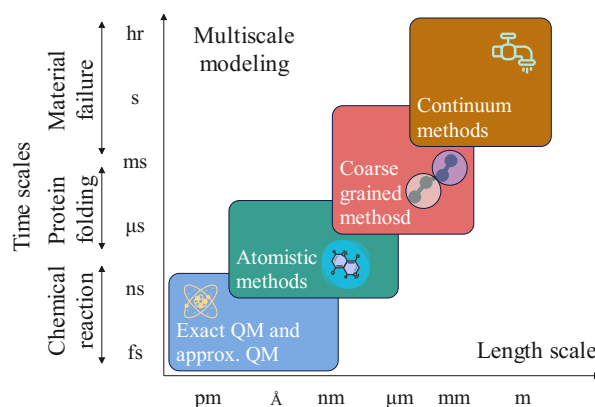


Fig. 2. Schematic diagram of multi-scale simulation hierarchy (Keith et al., 2021).

flooding agents under extreme conditions, providing a solid and reliable basis for the performance prediction of chemical flooding agents. Besides, this technology can continuously supply high-quality training data for generative molecular design, thereby ingeniously alleviating the dilemma faced by large models due to insufficient training data to a certain extent and paving an efficient and highly valuable research path for molecular design.

2.2 Intelligent molecular screening & design

Rapid screening and molecular generation technologies based on ML and DL have been used for fruitful explorations in fields such as drug development and protein structure prediction, accumulating a wealth of experience. These technological accumulations and experiences provide new ideas for the design of oilfield chemicals.

2.2.1 High-throughput screening of molecules

In the initial stage of molecular screening, quantitative structure-activity relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) models can utilize molecular fingerprints, topological indices or GNN to represent molecular structures, and combined with algorithms such as RF, SVM or CNN to quickly predict key properties (such as IFT, CMC, Hydrophilic-Lipophilic Balance (HLB) value and complex behaviors such as adsorption/wetting interactions between molecules and rock interfaces). This strategy significantly improves the efficiency of screening potential advantageous structures from a vast molecular library. If the study in question involves micelle assembly, oil-water interface modification, or molecular stability under high-temperature and high-salinity conditions, it can be combined with MD or DPD simulations to rapidly assess performance under different temperatures, salinities and pore environments.

In the field of predicting molecular structures and properties, the introduction of DL and GNN technologies has significantly propelled the paradigm shift of QSAR/QSPR models (for a comparison between traditional QSAR and deep QSAR, please refer to Table 2). In contrast to traditional representation methods that rely on manual feature engineering, such as molecular fingerprints and topological indices, the end-to-

Table 2. Comparison of traditional and deep QSAR models.

Dimension	Traditional QSAR	Deep QSAR
Input representation	Handcrafted molecular descriptors	Raw molecular formats
Feature engineering	Expert-defined descriptors	Automated feature learning
Model architecture	Linear/nonlinear statistical models	DNNs
Data requirements	Small datasets ($n < 10^3$); required high-quality features	Large datasets ($n > 10^4$)
Interpretability	High	Low
Computational cost	Low-cost CPU execution (minutes to hours)	GPU/TPU-accelerated training (hours to days)
Typical applications	Boiling point/solubility prediction; Early-stage ADMET screening	Virtual high-throughput screening

end training mechanism of graph-structured DL can automatically extract molecular graph embedding representations that contain complex structural information. This representation method not only breaks through the dimensional limitations of traditional methods in describing multi-scale molecular interactions but also achieves the synergistic optimization of atomic-level local features and molecular-level global features via graph convolution operations, thereby significantly improving the accuracy and generalization performance of the prediction model (Sippl et al., 2018; Tropsha et al., 2024). The successful application of a DL-based QSAR system using image representations for predicting agonists and antagonists is showcased in the work of Matsuzaka and Uesawa (2022). This system converts three-dimensional chemical structures into multi-angle two-dimensional images and uses convolutional neural networks for feature extraction, achieving an efficient and accurate classification prediction of compound activity. The optimized system maintains high prediction performance while reducing model training time and lowering experimental validation costs, providing important references for high-throughput screening and molecular innovation in practical applications.

In terms of model interpretability and computational efficiency, some novel regression strategies have shown unique advantages. For example, the “topological regression” method proposed by Zhang et al. (2024) innovatively constructs a sparse isometric mapping model between chemical space and activity space. By mathematically modeling topological invariants, it improves the model’s interpretability by more than 40% while maintaining prediction performance comparable to DL and reduces the computational time by about two orders of magnitude. This modeling strategy, which combines prediction efficiency with mechanistic interpretability, provides an efficient computational tool for the rapid identification of lead compounds in high-throughput screening and, at the same time, establishes a theoretical framework for structure-based molecular optimization.

Current high-throughput screening strategies are gradually evolving towards data-driven and multi-level collaborative optimization approaches. By integrating experimental data, simulation results and field parameters, intelligent QSAR/QSPR platforms can quickly screen out ideal candidate

molecules from large-scale molecular libraries and conduct multi-objective evaluations of these molecules at the same time. For example, while predicting key parameters such as IFT, CMC and HLB value, the platform can also consider the adsorption and wettability behavior between molecules and rock interfaces, providing comprehensive and precise guidance for the formulation design of oilfield chemicals. As data volumes continue to grow and computing resources are further improved, the fusion of multi-modal data (including structural images, molecular descriptors and dynamic simulation data) is becoming the mainstream trend in high-throughput screening. This will help to build higher-precision prediction models and also promote the transition from virtual screening to generative molecular design, providing a more solid technical foundation for the development and field application of new oilfield chemicals.

2.2.2 Generative AI-driven chemical agent design

Building on the traditional supervised learning approach for constructing performance prediction and molecular property association models, the novel method of Generative Molecular Design (GMD) has shown rapid evolution in recent years. This method can intelligently explore and “generate” new molecules in high-dimensional chemical space by learning existing molecular structures and their performance data through deep neural networks (Grantham et al., 2022; Nnadili et al., 2023; Yao et al., 2023; Du et al., 2024). Its core advantage lies in utilizing architectures such as Generative Adversarial Networks (GANs) (Kadurin et al., 2017; Mao et al., 2020; Liu et al., 2023), VAE (McLoughlin et al., 2023; Zhou and Huang, 2024), and Diffusion Models (Abramson et al., 2024) to learn underlying patterns from vast molecular data and generate novel molecular structures with specific functions (Bhowmik et al., 2024; Chen et al., 2024; Nnadili et al., 2024; Nguyen and Karolak, 2025), as shown in Fig. 3. Compared to traditional “trial-and-error” or purely random search methods, generative AI is superior at identifying potential structural patterns and recombining or modifying molecules accordingly.

For generating molecules with specific properties, Kadurin et al. (2017) proposed the druGAN model, which combines GAN with Autoencoders for new molecule generation and optimization. The focus of their study is on directed generation

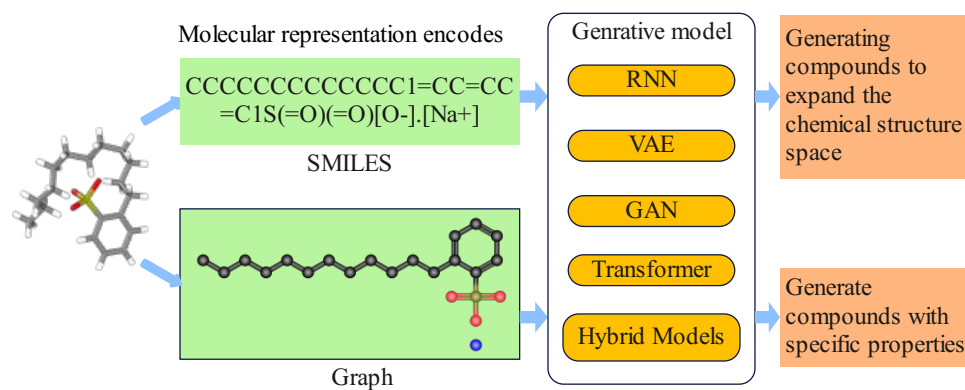


Fig. 3. Workflow of generative AI for molecular design.

capabilities, enabling AI to generate molecules with preset anti-cancer properties and to handle large-scale molecular databases. McLoughlin et al. (2023) introduced the ATOM-GMD platform, which integrates VAE with genetic algorithms to generate a large number of molecular candidates in batches. The top 103 compounds were experimentally synthesized and tested, verifying the potential of this platform in multi-parameter optimization for drug development.

Targeting specific application scenarios, Nnadili et al. (2024) proposed an AI-driven molecular design framework specifically for surfactants, integrating generative models (VAE), predictive modeling (GNN), and Reinforcement Learning (RL). This framework uses self-referencing embedded strings (Krenn et al., 2020) as the molecular representation method to ensure the validity of generated molecules and employs VAE for molecular structure generation and GNN for target property prediction. Furthermore, reinforcement learning optimizes the generated molecules in the latent space of VAE to enhance their performance in target properties. The results indicate that this framework can generate novel surfactant molecules with stable structures and chemical rules, which not only meet the target CMC threshold but also demonstrate higher thermodynamic stability. The structural diversity of these molecules is assessed through similarity analysis, and their stability and solvation free energy are verified through molecular dynamics simulations. This study provides a new concept for the intelligent design of surfactant molecules, which can be applied in the future to EOR, detergents and emulsifiers, demonstrating the full potential of AI to perform molecular optimization.

QSAR/QSPR models rapidly predict IFT, CMC and other properties, and can be combined with molecular dynamics simulations to evaluate performance under high-temperature and high-salinity conditions. DL and graph convolution methods significantly enhance prediction accuracy and interpretability. Generative AI can “create” functional molecules in high-dimensional chemical space and optimize their key properties through techniques such as reinforcement learning, targeting the complex and diverse application environments of oilfields to facilitate the efficient and rapid design of “targeted” functional oilfield chemicals. The fusion of multi-modal data and multi-objective evaluation endows intelligent molecular design

with great potential, accelerating the transition from virtual screening to GMD and bringing innovational ideas in the field of oilfield chemicals.

2.3 Chemical agent formulation optimization and performance prediction

In the critical transition stage from the laboratory to the field application of oilfield chemicals, “formulation optimization” and “performance prediction” often determine whether EOR technologies can achieve the expected results and directly relate to their economic feasibility and promotion value. Current research in the field of oilfield chemicals is undergoing a paradigm shift from traditional empirical trial-and-error to multi-objective intelligent optimization design, which has significant potential for future development.

2.3.1 Interface regulation and performance prediction

One of the core aims of chemical flooding is to reduce oil-water or oil-gas IFT, in order to promote crude oil emulsification, and to modify rock wettability, in order to reduce capillary resistance for imbibition oil recovery, thereby enhancing oil and gas recovery rates (Keradeh and Khanghah, 2024; Saberi et al., 2024; Yousefmarzi et al., 2024). Mouallem et al. (2024) predicted CO₂-brine IFT using gradient boosting models to optimize deep storage strategies for CO₂ in carbonate aquifers in the UAE. Yousefmarzi et al. (2024) systematically compared the performance of six algorithms (Support Vector Regression (SVR), RF, etc.) in predicting IFT in oil/gas and oil/water systems, with SVR and CatBoost achieving prediction accuracies of $R^2 = 0.99$ for oil/gas systems. Rashidi-Khaniabadi et al. (2023) modeled surfactant-hydrocarbon IFT using integrated tree-based ML algorithms (Decision Tree (DT), ET, GBRT), finding that GBRT achieved the best prediction accuracy. Their study provides an efficient prediction tool for EOR in petroleum engineering. Other scholars combined DL with decision trees to compare the contributions of different salt molecular weights, temperatures, and oil properties to IFT, offering more specific guidance for optimizing additive formulations under field conditions. Regarding complex salt environments, Liu et al. (2024) developed an IFT prediction model for oil-water systems based on real crude oil samples and salt types. Using a DT model,

Table 3. Performance of ML algorithms for training and testing data in % oil recovery of CO₂ foam EOR (Iskandarov et al., 2024a).

Algorithms	R ²		MAE		RMSE		WAPE	
	Train	Test	Train	Test	Train	Test	Train	Test
DT	1	0.95	0	1.51	0	1.89	0	5.47
RF	0.99	0.98	0.39	1.04	0.59	1.48	1.22	3.23
Gradient boosting	0.99	0.99	0.14	1.06	0.17	1.54	0.44	3.31
Extreme gradient boosting	0.99	0.98	0.59	1.25	0.82	1.67	1.84	3.91
Extremely randomized trees	1	0.99	0.01	0.97	0.02	1.32	0.04	3.05
DNN	0.99	0.99	0.36	0.64	0.47	1.01	1.18	2.31

they revealed that temperature had a weight of 46.3% on IFT, significantly higher than salt (8.7%) and other factors.

Another key mechanism in chemical flooding is rock wettability regulation. Ibrahim (2023) proposed a new efficient pathway for predicting shale wettability using ML methods. They collected 250 sets of contact angle experimental data covering different shale types and experimental conditions (such as temperature, pressure, salinity, etc.) and constructed various ML models, including Linear Regression, DT, RF, Functional Network (FN), and GBRT, to predict contact angles in CO₂-water-shale ternary systems. The results showed that non-linear ML models could more accurately fit the complex relationship between input parameters and wettability than traditional linear regression models. The GBRT model performed the best, with determination coefficients (R²) of 0.99 and 0.98 for training and testing sets, respectively, and a Root-Mean-Square Error (RMSE) of less than 5 degrees. Sensitivity analysis further indicated that pressure is the key factor affecting shale wettability, with shale exhibiting strong water wettability at low pressures and CO₂ wettability at high pressures. This study significantly improved the efficiency and accuracy of shale wettability prediction, building a strong technology foundation for CO₂ geological storage site selection, sealing assessment, and oil and gas recovery enhancement. Keradeh and Khanghah (2024) systematically studied the application potential of diethylenetriaminepentaacetic acid (DTPA) in sandstone wettability modification and found that DTPA could significantly change rock wettability from oil-wet to strongly water-wet. They revealed that a concentration of 5 wt% DTPA achieved the best wettability modification effect, while the presence of key determining ions (PDIs, such as Ca²⁺, Mg²⁺, and SO₄²⁻) at three times the original concentration weakened this effect. In addition, they combined RF and Boosted Regression Tree ML models to predict contact angles from 240 experimental datasets, and the Boosted Regression Tree model demonstrated superior prediction capabilities (R² > 0.999). Sensitivity analysis indicated that the main factors affecting wettability were PDIs, salinity, reaction time, and DTPA concentration. This work not only validated the feasibility of DTPA in regulating rock wettability under high-salinity conditions but also showcased the application potential

of ML in predicting and optimizing flooding systems.

2.3.2 Displacement system optimization and performance prediction

Foam and microemulsion flooding play important roles in enhancing sweep efficiency and improving flow control (Kamaludin et al., 2024; Maia et al., 2024). The multi-component synergistic effects, non-linear flow characteristics and complex interfacial behaviors of these systems often make it difficult to conduct large-scale screening using experimental methods alone. Nonetheless, ML models offer a pathway for rapid prediction and combinatorial optimization.

Regarding CO₂ foam systems, Iskandarov et al. (2024a) utilized ML techniques to predict CO₂ foam performance based on key parameters such as foam apparent viscosity and IFT. They employed six different models (with comparative results shown in Table 3) and found that DNN performed exceptionally well in predicting oil recovery rates in reservoirs. This is because in the test set predictions, the DNN model achieved the lowest values for Mean Absolute Error (MAE) at 0.64, RMSE at 1.01, and Weighted Absolute Percentage Error (WAPE) at 2.31%. The findings revealed that foam apparent viscosity and IFT are the key factors affecting recovery rates. Appropriately increasing foam viscosity and reducing IFT can significantly enhance recovery rates, while the effect plateaus after a certain threshold. Iskandarov et al. (2024b) also explored the impact of different surfactant types and operating conditions on CO₂ foam performance. Using ML models to predict foam apparent viscosity, they found that nonionic and cationic surfactants exhibited better tolerance to high salinity conditions. Moreover, the HLB value of surfactants significantly influenced foam strength. The aforementioned studies provide important theoretical bases and practical guidance for optimizing CO₂ foam for EOR and carbon sequestration technologies.

The EACN is a key parameter for measuring the hydrophobicity of oil compounds and as such is significant in the design of surfactant/oil/water systems and microemulsion applications (Chang et al., 2019; Qu et al., 2023). Traditional experimental methods for determining EACN are cumbersome, time-consuming and require high sample purity (Wan et al., 2016).

In recent years, ML techniques have provided new ways for rapid and accurate EACN prediction. Delforce et al. (2022) developed models based on Graph Machines (GM) and Neural Networks (NN) to quickly predict the EACN of oils. They used a database of 121 compounds, with GM predicting from SMILES codes and NN predicting from COSMO-RS calculated σ -moments. They found that both GM and NN models performed comparably in terms of prediction accuracy, but in the case of homologous series prediction, the GM model showed better consistency with the experimental results. Furth et al. (2024) compared the performance of three GNNs and an XGBoost model in EACN prediction. They collected EACN data for 183 organic molecules, converted molecular structures into SMILES codes, and studied the impact of geometric optimization on prediction. The results indicated that the Crystal Graph Convolutional Neural Network model trained with MMFF94 optimized geometric data performed the best, with a prediction error of 1.15 EACN units and an R^2 score of 0.9. Meanwhile, the XGBoost model excelled in terms of runtime and prediction accuracy, especially for small datasets.

Both of the above studies demonstrated that ML models can rapidly and accurately predict EACN, hence are powerful tools in the design of microemulsion systems, significantly reducing experimental workload and time costs.

2.3.3 Composite drive design and performance prediction

Polymer flooding is an important technology in tertiary oil recovery, with the viscosity of polymer solutions directly affecting sweep efficiency (Dai et al., 2023). Utilizing Regression Decision Trees, SVR, and Multi-Layer Perceptrons, the viscosity of modified partially Hydrolyzed Polyacrylamide solutions can be accurately predicted under high salinity and different shear rates and temperature scenarios (Rashidi-Khaniabadi et al., 2023; Shakeel et al., 2023). Shakeel et al. (2023) established a viscosity model using Artificial Neural Network in high-salinity environments (up to 167,000 ppm), whose training and testing set correlation coefficients (R^2) both exceeded 0.99, significantly enhancing the efficiency and reliability of field formulation design.

Surfactant-Polymer (SP) composite flooding requires the optimization of both technical and economic indicators. Larestani et al. (2022) developed a cascade neural network model that excelled in the joint prediction of Recovery Factor and Net Present Value (NPV), with mean absolute errors of 0.66% and 1.95%, respectively. Sun et al. (2021) combined the hydrophilic-lipophilic difference - net average curvature equation with a neural network surrogate model to construct a techno-economic evaluation framework for alkali/surfactant/polymer flooding. The neural network was used to replace the numerical simulator to predict oil and water dynamics, combined with particle swarm optimization and Pareto optimality to reveal the trade-off relationship between NPV and Chemical Efficiency (CE). When NPV reached \$933,000, CE was \$8.21 per barrel; when CE was reduced to \$7 per barrel, NPV only decreased by 3.3%. Monte Carlo analysis indicated that the project would incur losses if oil prices fell below \$30 per barrel. Multi-objective optimization quantified the

association between water cut and economic benefits, offering risk decision support for alkali/surfactant/polymer schemes.

The general research concept for formulation optimization and performance prediction is shifting from traditional “empirical summary + trial-and-error experimentation” to “data-driven + multi-objective optimization”. By integrating multi-source experimental data and field operation information, ML models can rapidly assess key indicators such as interfacial tension, foam stability, polymer viscosity, and economic feasibility within a short period. With the development of digital oilfields, multi-disciplinary collaboration and high-throughput experimental platforms, the value of AI in formulation optimization for oilfield chemicals will continue to expand, providing more robust data support and innovative ideas for further enhancing the oil recovery rates and extending the economic lifespan of aging oil fields.

2.4 Intelligent experimental systems

In the process of material development, as the advancement of molecular design and simulation technologies continues, the workload and parameter combinations required for experimental verification have also increased significantly. Specifically, the high dimensionality of data makes it difficult for traditional manual operation modes to accurately capture the required information. With the emergence of high-throughput intelligent experimental systems that integrate automation, robotics and intelligent algorithms, the entire “design - synthesis - testing - analysis” process is interconnected, enabling unmanned or minimally manned autonomous research processes. While the molecular structure designed by AI is synthesized automatically through the robot chemist platform, the synthetic product is detected in real time through online detection, and the detection results are fed back to the AI model for optimization, thus achieving rapid iteration. This strategy not only shortens the material development cycle but also makes the research process more sustainable (Rahmanian et al., 2022; Ha et al., 2023; Sadeghi et al., 2024; Tom et al., 2024).

With respect to intelligent systems, researchers have made many fruitful attempts and achieved gratifying progress in recent years. Seifrid et al. (2022) developed the Self-Driving Lab platform ChemOS (Roch et al., 2018), which combines automated synthesis, high-throughput characterization (HPLC-MS, optical analysis), and Bayesian optimization algorithms (Phoenics, Gryffin, etc.), to achieve the efficient design and optimization of Organic Semiconductor Laser materials and inorganic thin-film materials. Through the closed-loop “design-manufacture-test-analyze” process, the material development cycle was shortened from months to days, generating high-quality, shareable datasets. This platform promotes the automation, reproducibility and interdisciplinary collaboration of chemical discoveries and acts an accelerated tool for addressing global challenges such as energy and environment. Szymanski et al. (2023) developed the A-Lab autonomous laboratory, which integrates computational screening (Materials Project and Google DeepMind stability data), literature mining ML models (NLP synthesis recipe recommendations), and robotic experimental systems, as shown in Fig. 4. Within

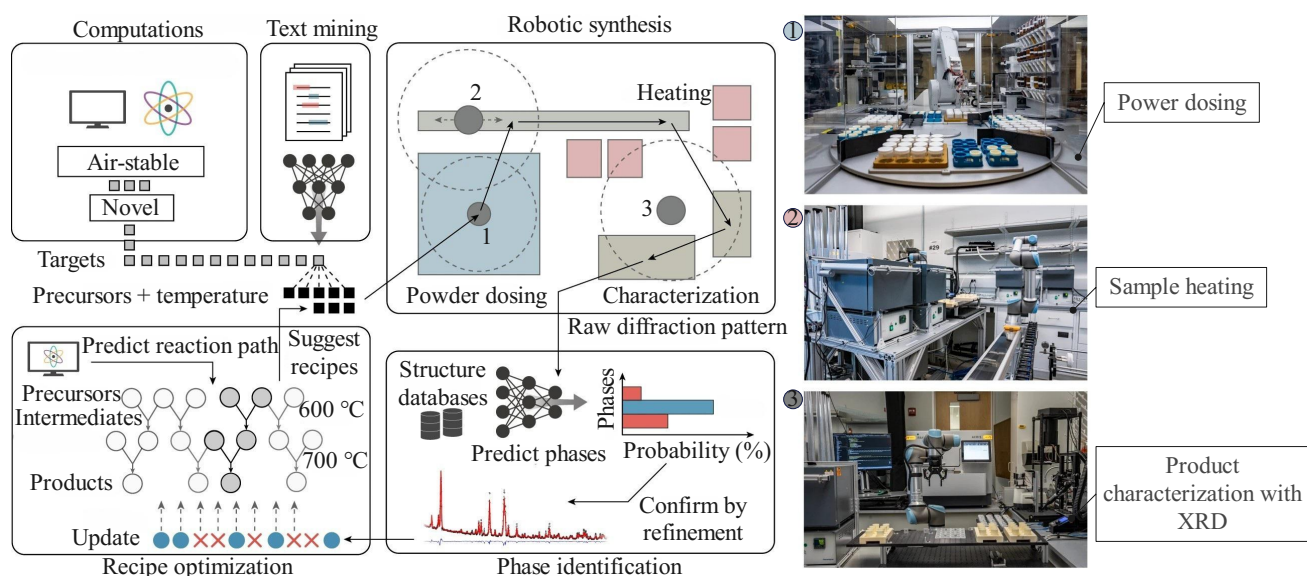


Fig. 4. Architecture of the A-Lab autonomous materials discovery platform (Szymanski et al., 2023).

Table 4. Core architecture of the iChemFoundry platform (Lu et al., 2024).

Module	Function	Key technologies
Automated synthesis system	Parallel execution of hundreds to thousands of reaction conditions.	Flow chemistry workstations Microplate-based batch reaction systems Robotic manipulator platforms
High-throughput detection system	Real-time analysis of products within seconds to minutes.	Online UPLC-MS Microfluidic IR/UV spectroscopy MALDI-TOF mass spectrometry
AI decision-making system	Optimizes reaction conditions, designs synthetic pathways, generates novel molecular structures, and dynamically adjusts experimental plans.	Machine learning models Retrosynthesis algorithm libraries Generative molecular design engines

merely 17 days, they successfully synthesized 41 new inorganic materials (oxides/phosphates, etc.) and verified the synthesizability of 71% of the computationally predicted materials. This study is the first to realize the full closed-loop autonomous optimization of solid-state material synthesis, proving the feasibility of AI-driven platforms in accelerating material discovery and providing a practical example for the computational-experimental collaborative material development paradigm.

The intelligent chemical synthesis platform iChemFoundry, developed by Zhejiang University (Lu et al., 2024), integrates automated high-throughput experimental systems with AI decision-making systems to achieve the full closed-loop optimization of “design-experiment-feedback”. Table 4 lists the core architecture of the iChemFoundry platform. This platform, with its μL -scale microreactor system, significantly reduces reagent consumption (1/100 to 1/1000 of traditional methods) and supports solid/liquid phase reactions, photo/electrocatalysis, and other complex systems. It also employs online

ultra performance liquid chromatography-mass spectrometry, and MALDI-TOF mass spectrometry for real-time product analysis (yield, purity, byproducts) within seconds to minutes. To illustrate the efficiency of this platform, through five rounds of AI optimization (1,200 experiments), the enantioselectivity was increased from 68% to 95%; in just 4 days, it constructed a library of over 200 anti-tumor derivatives, with an efficiency five times higher than traditional methods. In multi-step serial synthesis, the total product yield was successfully increased from 5% to 22%, demonstrating the platform’s ability to control complex systems. Compared with similar platforms (such as Chemputer, IBM RXN), iChemFoundry features the advantage of deep collaboration between high-throughput experiments and intelligent algorithms, promoting the transition of chemical synthesis from experience-driven to data-driven paradigms, providing an efficient and green solution for drug development and functional molecular design. In the future, through modular expansion and integration with quantum computing, this platform is expected to further break through

the bottleneck of reaction mechanism analysis and cross-scale synthesis.

Currently, intelligent experimental systems are driving the transition of R&D towards high-throughput, automated, and intelligent directions (Tom et al., 2024). If applied to oilfield chemical research, the deep integration of robotic chemists, high-throughput screening platforms, self-driving laboratories, and knowledge graph technologies can not only test a wider range of molecular or formulation combinations in a shorter period but also achieve dynamic balance and global optimization of different environmental variables and multiple performance indicators. This could enable the screening of formulations adapted to harsh conditions such as high salinity and temperature at lower costs and shorter cycles.

3. Technical challenges & future directions

Recently, AI has shown great potential in several links of new material development, including molecular simulation, data screening and performance prediction, formulation optimization, and intelligent experimental verification. However, regarding the complex and harsh oilfield environment, realizing the implementation and promotion of this technology in oilfield chemicals still faces many challenges; how to break the deadlock and find development opportunities is a key issue worth considering in the field of oilfield chemicals.

3.1 Technical challenges

3.1.1 Data quality and sample scarcity

Oilfield chemicals often need to adapt to extreme conditions such as high temperature, salinity and pressure. The experimental data acquisition process is costly and time-consuming, and the data is often “fragmented” and dispersed across different laboratories or fields (Li et al., 2021; Waqar et al., 2023). Under such circumstances, the available datasets are not only relatively limited in scale but also have differences in data generation conditions, making data standardization difficult. At the same time, the extrapolation capability of AI models outside the range of training data (i.e., their ability to predict accurately in conditions not covered by the training data) remains an imminent problem. Once extrapolated to more severe conditions or other oilfield blocks, the model may fail due to the lack of sufficient training samples. In the face of this “data scarcity”, it has become urgent to build a high-quality oilfield chemicals database and combine transfer learning, few-shot learning, active learning, and other strategies to enhance the robustness and extrapolation capability of the models.

3.1.2 Multi-scale mechanistic coupling and model complexity

Oilfield chemical reaction processes differ at multiple scales including time and space: from the nanosecond-level dynamics at the molecular level to the hour-level or day-level macroscopic flow in the reservoir, and even to the monthly or yearly dimensions of reservoir engineering (Peter and Kremer, 2009; Joshi and Deshmukh, 2021; Peng et al., 2021). Although multi-scale simulation has made great progress, there

is still a lack of a mature “adaptive coupling” solution that can seamlessly connect the fine mechanisms at the molecular level with pore network/macroscale flow models. For example, MLPs require a large amount of high-precision quantum chemical data for support, which has high computational requirements, while coarse-grained simulations may sacrifice the local electronic structure or molecular interaction details while expanding spatial scales. Therefore, how to balance computational efficiency and precision remains the biggest pain point in multi-scale research.

3.1.3 Algorithm efficiency and real-time response conflict

High-precision DL models or generative AI often require a large amount of training data and involve an expensive training process. However, in the area of oilfield production, production data changes in real time and decisions need to be made quickly and robustly, without waiting for lengthy offline calculations. A key direction of concern for both industry and academia is finding the means to use model pruning, mixed precision computing, edge computing, and other methods to embed AI models into the field links, in order to achieve quasi-real-time or online parameter optimization and formula updates. Besides, computational power limitations in some oilfields cannot be ignored. Although large-scale cloud computing can be feasible, its network bandwidth and reliability may also have a certain impact on real-time prediction.

3.1.4 Experimental system and field data closed-loop insufficiency

Although “self-driving laboratories” and “high-throughput automated platforms” have been well-developed in the materials and pharmaceutical industries (Fakhruldeen et al., 2022; Szymanski et al., 2023; Lu et al., 2024), adapting such platforms to oilfield chemical research still needs to deal with more complex conditions such as high-salinity, high-temperature, high-pressure and handle multi-type multi-source data (rock mineral composition, salinity, viscosity, IFT, etc.). In addition, field data is often burdened with uncontrollable factors, such as sampling errors, instrument instability, and well condition differences, potentially leading to a “disconnect” between laboratory verification results and actual field needs. If there is no efficient data feedback mechanism to update the model, a true “closed-loop” R&D system cannot be formed.

3.1.5 Industrialization and cost-benefit trade-offs

AI technology performs well at the experimental stage, but in large-scale applications, it is necessary to balance the costs, risks and long-term benefits involved. In addition, the oilfield development environment is highly variable, and the deployment of AI-driven automated devices or sensor networks requires a large initial investment and high technical barriers. Also, many oilfield companies have a high degree of confidentiality for core data, making it difficult to share or centrally train models, which also adds to the challenges faced by large-scale data-driven algorithms. Therefore, how to achieve a balance between economic and technical feasibility is a problem faced by all parties.

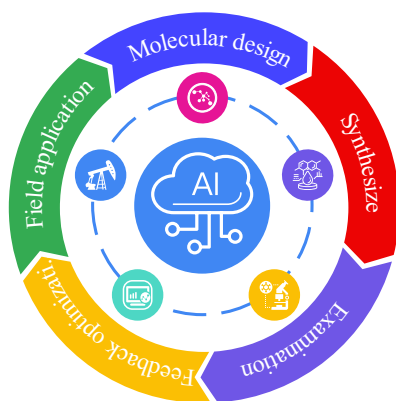


Fig. 5. Paradigm of oilfield chemical development empowered by AI.

3.2 Future trends

3.2.1 Intelligent laboratories and data acquisition

In the future, the seamless integration of automated synthesis platforms, online monitoring devices, and AI algorithms are expected to play a pivotal role in enhancing the efficiency and quality of data acquisition (O'Neill, 2021; Seifrid et al., 2022; Lu et al., 2024). First, the application of high-throughput intelligent experimental systems can rapidly generate high-quality standardized data via automated synthesis platforms and online monitoring devices, effectively alleviating the problem of data scarcity. Second, data augmentation and transfer learning techniques can migrate data from other fields or similar oilfield environments to the scenarios of the target oilfield, thereby enhancing the applicability of the models. Additionally, active learning strategies can prioritize the collection of the most valuable samples for model training, reducing the blindness of data collection and improving the data utilization efficiency. Finally, by establishing cross-laboratory and cross-oilfield data-sharing platforms, dispersed data resources can be integrated into a unified standardized database, providing more comprehensive training data for AI models. These measures not only can resolve the issues of data scarcity and standardization but also offer more efficient data support for the future development of oilfield chemicals.

3.2.2 Multi-scale models and Physics-Informed Neural Networks (PINNs)

To better integrate physical mechanisms with data-driven approaches, PINNs and multi-fidelity learning methods are becoming the current research hotspots. To improve the extrapolation capabilities of AI models under extreme conditions, certain improvements can be made. First, embedding the physical laws of oilfield chemical reactions (such as thermodynamic constraints and fluid dynamic equations) directly into the structure of PINNs models can significantly enhance their predictive capabilities under extreme conditions. Second, multi-fidelity learning methods combine high-precision quantum chemical data with large-scale, lower-precision simulation data to effectively improve model generalization. Third, dynamic data feedback mechanisms can collect data in real time

during experiments and field deployments and feed it back into the models. Through online optimization algorithms, model parameters can be dynamically adjusted to enhance model adaptability. By employing these methods, more accurate predictions and more efficient model optimization can be achieved in complex oilfield environments, providing enhanced technical support for the development of oilfield chemicals.

3.2.3 Multi-objective optimization and economic-environmental co-evaluation

Future oilfield development will undoubtedly place a greater emphasis on environmental protection and sustainability. AI can incorporate multiple indicators such as interfacial viscosity reduction, oil recovery efficiency, environmental risk, and economic returns into a single framework for comprehensive assessment. As the concept of green chemistry deepens, the development of oilfield chemical agents will shift from a single pursuit of “high oil recovery efficiency” to a multi-dimensional goal of “low environmental risk, biodegradability, and low carbon emissions”. By using AI algorithms to find “Pareto optimal solutions”, companies can make scientifically-based decisions under the dual pressures of policy and market.

3.2.4 Algorithm simplification and field-deployable technologies

Even when data and models achieve excellent results in the laboratory, real oilfield sites continue to face the challenges of computational power limitations, network latency, and hardware compatibility. Future efforts can focus on model pruning and knowledge distillation to construct lightweight inference models that can run stably on on-site industrial computers or edge devices. For more complex computational needs, hybrid cloud or distributed computing architectures can be implemented, with some tasks transferred to remote data centers with on-site systems retaining only the essential fast-response modules, achieving “cloud-edge collaboration”.

3.2.5 Full-process closed-loop and digital twin oilfields

With the continued improvement of data platforms and automation equipment, the development of oilfield chemicals can hope to form a complete closed loop of “design-synthesis-testing-feedback-field deployment”, as illustrated in Fig. 5. Within the framework of digital twin oilfields, virtual models and real oilfields can interact in real time through data, combining online optimization algorithms to refine predictions of displacement processes and automatically adjust injection strategies or formulation compositions when necessary, thus achieving a proactive control of oilfield production. This will greatly enhance development efficiency and visualization levels and promote the transition of oilfield production towards intelligent, low-carbon and efficient operation and maintenance directions.

Overall, AI has broad prospects regarding the R&D and application of oilfield chemicals, but breakthroughs are still needed in data, mechanism coupling, and on-site feasibility. In the future, the innovative support role of AI for the next generation of oilfield chemicals can truly be realized through high-throughput intelligent experimental systems to obtain

more complete and accurate multi-source data, relying on multi-scale simulations and physics-informed learning models to build cross-scale, multi-objective optimization frameworks, and combining digital twin technologies to achieve seamless integration between experiments and production. On this basis, greening and sustainable development will inject new research momentum into the field, further expanding the application boundaries of AI-driven oilfield chemicals.

4. Conclusions

The integration of AI into the R&D of oilfield chemicals heralds a new era of innovation and efficiency. The transformative impact of AI and outline future directions for this field can be highlighted in the following key points:

- 1) **Paradigm Shift in Research Methodology:** AI-driven approaches are revolutionizing oilfield chemical development by replacing traditional trial-and-error methods. Through DL and pattern recognition, AI can predict the molecular properties and screen for novel chemical formulations tailored to specific reservoir conditions, significantly reducing R&D cycles and associated costs.
- 2) **Enhanced Understanding via Multi-scale Simulation:** MLPs and enhanced sampling techniques are providing unprecedented insights into the behavior of flooding agents under extreme conditions. These tools enable the analysis of complex molecular interactions across multiple scales, from nanosecond dynamics at the molecular level to hour-scale displacements in reservoirs, enriching our comprehension of intricate mechanisms.
- 3) **Advancements in Intelligent Experimental Systems:** The emergence of high-throughput intelligent experimental platforms is accelerating the entire development process. These systems integrate automated synthesis, real-time detection, and AI-driven decision-making, allowing for the rapid iteration and optimization of chemical formulations. This integrative approach not only shortens the development timeline but also enhances the sustainability and efficiency of the research process.
- 4) **Challenges and Opportunities:** Despite the significant progress, challenges persist in data quality, model interpretability and on-site deployment. Most importantly, the scarcity of high-quality, standardized data and the computational demands of complex AI models are hurdles that require innovative solutions. However, with the advancement of cloud computing and digital twin technologies, the vast potential of AI in this field can still be realized.
- 5) **Future Directions:** Looking ahead, the development of oilfield chemicals will increasingly emphasize environmental protection and sustainability. AI will continue to play a pivotal role in multi-objective optimization, balancing oil recovery efficiency with environmental risk and economic returns. In addition, the simplification of algorithms and the development of field-deployable technologies will be crucial for broader adoption in the industry. The vision of a fully closed-loop system, from design to field deployment, supported by digital twin

oilfields, promises to enhance development efficiency and drive the industry toward intelligent, low-carbon operations.

In conclusion, AI is not merely a supplementary tool but a fundamental driver of innovation in oilfield chemical development. As technology continues to evolve, its integration into the industry will likely yield more efficient, environmentally friendly and economically viable solutions for EOR.

Acknowledgements

This work was supported by the CNPC Key Science and Technology Special Project “Research on Novel Methods and Technologies for Enhanced Oil Recovery” (No. 2023ZZ04).

Additional information: Author’s email

lweifeng@petrochina.com.cn (W. Lyu); pmg@cczu.edu.cn (M. Peng).

Conflict of interest

The authors declare no competing interest.

Open Access This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

- Abramson, J., Adler, J., Dunger, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024, 630(8016): 493-500.
- Aldossary, A., Campos-Gonzalez-Angulo, J. A., Pablo-Garcia, S., et al. In silico chemical experiments in the age of AI: From quantum chemistry to machine learning and back. *Advanced Materials*, 2024, 36(30): 2402369.
- Batzner, S., Musaelian, A., Sun, L., et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 2022, 13(1): 2453.
- Bertazzo, M., Gobbo, D., Decherchi, S., et al. Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy. *Journal of Chemical Theory and Computation*, 2021, 17(8): 5287-5300.
- Bhowmik, D., Zhang, P., Fox, Z., et al. Enhancing molecular design efficiency: Uniting language models and generative networks with genetic algorithms. *Patterns*, 2024, 5(4): 100947.
- Chang, L., Pope, G. A., Jang, S. H., et al. Prediction of microemulsion phase behavior from surfactant and co-solvent structures. *Fuel*, 2019, 237: 494-514.
- Chen, K., Li, J., Wang, K., et al., Chemist-X: Large language model-empowered agent for reaction condition recommendation in chemical synthesis. *ArXiv Preprint ArXiv: 2311.10776*, 2024.
- Dai, C., You, Q., Zhao, M., et al. *Principles of Enhanced Oil Recovery*. Singapore, Springer Nature Singapore, 2023.
- Delforce, L., Duprat, F., Ploix, J. L., et al. Fast prediction

- of the equivalent alkane carbon number using graph machines and neural networks. *ACS Omega*, 2022, 7(43): 38869-38881.
- Druetta, P., Raffa, P., Picchioni, F. Chemical enhanced oil recovery and the role of chemical product design. *Applied Energy*, 2019, 252: 113480.
- Du, Y., Jamasb, A. R., Guo, J., et al. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 2024, 6(6): 589-604.
- Fakhruldeen, H., Pizzuto, G., Glowacki, J., et al. AR-Chemist: Autonomous robotic chemistry system architecture. Paper Presented at 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, 23-27 May, 2022.
- Furth, N. R., Imel, A. E., Zawodzinski, T. A. Comparison of machine learning approaches for prediction of the equivalent alkane carbon number for microemulsions based on molecular properties. *The Journal of Physical Chemistry A*, 2024, 128(32): 6763-6773.
- Grantham, K., Mukaidaisi, M., Ooi, H. K., et al. Deep evolutionary learning for molecular design. *IEEE Computational Intelligence Magazine*, 2022, 17(2): 14-28.
- Ha, T., Lee, D., Kwon, Y., et al. AI-driven robotic chemist for autonomous synthesis of organic molecules. *Science Advances*, 2023, 9(44): eadj0461.
- Horstemeyer, M. F. Multiscale modeling: A review, in *Practical Aspects of Computational Chemistry: Methods, Concepts and Applications*, edited by J. Leszczynski and M. K. Shukla, Springer Netherlands, Dordrecht, pp. 87-135, 2010.
- Ibrahim, A. F. Prediction of shale wettability using different machine learning techniques for the application of CO₂ sequestration. *International Journal of Coal Geology*, 2023, 276: 104318.
- IEA. *World Energy Outlook 2024*, IEA, Paris, France, 2024.
- Iskandarov, J., Ahmed, S., Fanourgakis, G. S., et al. Predicting and optimizing CO₂ foam performance for enhanced oil recovery: A machine learning approach to foam formulation focusing on apparent viscosity and interfacial tension. *Marine and Petroleum Geology*, 2024a, 170: 107108.
- Iskandarov, J., Fanourgakis, G. S., Ahmed, S., et al. Machine learning prediction and optimization of CO₂ foam performance for enhanced oil recovery and carbon sequestration: Effect of surfactant type and operating conditions. *Geoenergy Science and Engineering*, 2024b, 240: 213064.
- Joshi, S. Y., Deshmukh, S. A. A review of advancements in coarse-grained molecular dynamics simulations. *Molecular Simulation*, 2021, 47(10-11): 786-803.
- Kadurin, A., Nikolenko, S., Khrabrov, K., et al. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, 2017, 14(9): 3098-3104.
- Kamal, M. S., Hussein, I. A., Sultan, A. S. Review on surfactant flooding: Phase behavior, retention, IFT, and field applications. *Energy & Fuels*, 2017, 31(8): 7701-7720.
- Kamaludin, N. A., Suhaidi, N. N. S., Ismail, N. Green surfactants for enhanced oil recovery: A review. *Materials Today: Proceedings*, 2024, 107: 243-248.
- Kang, P., Shang, C., Liu, Z., Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. *Accounts of Chemical Research*, 2020, 53(10): 2119-2129.
- Karimov, D., Toktarbay, Z. Enhanced oil recovery: Techniques, strategies, and advances. *ES Materials & Manufacturing*, 2023, 23(2): 1005.
- Karniadakis, G., Beşkök, A., Aluru, N. *Microflows and Nanoflows: Fundamentals and Simulation*. New York, NY, Springer, 2005.
- Keith, J. A., Vassilev-Galindo, V., Cheng, B., et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical Reviews*, 2021, 121(16): 9816-9872.
- Keradeh, M. P., Khanghah, A. M. Experimental investigation and machine learning modeling of diethylenetriamine-pentaacetic acid agents in sandstone rock wettability alteration: Implications for enhanced oil recovery processes. *Journal of Molecular Liquids*, 2024, 404: 124959.
- Kirch, A., Razmara, N., Mamani, V. F. S., et al. Multiscale molecular modeling applied to the upstream oil & gas industry challenges. *Polytechnica*, 2020, 3(1): 54-65.
- Krenn, M., Häse, F., Nigam, A., et al. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 2020, 1(4): 045024.
- Laio, A., Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 2002, 99(20): 12562-12566.
- Larestani, A., Mousavi, S. P., Hadavimoghaddam, F., et al. Predicting the surfactant-polymer flooding performance in chemical enhanced oil recovery: Cascade neural network and gradient boosting decision tree. *Alexandria Engineering Journal*, 2022, 61(10): 7715-7731.
- Li, H., Yu, H., Cao, N., et al. Applications of artificial intelligence in oil and gas development. *Archives of Computational Methods in Engineering*, 2021, 28(3): 937-949.
- Liu, C., Wang, J., Wang, J., et al. Accurate modeling of crude oil and brine interfacial tension via robust machine learning approaches. *Scientific Reports*, 2024, 14(1): 28800.
- Liu, D., Zhang, F., Liu, Z., et al. A review of machine learning potentials and their applications to molecular simulation. *CIESC Journal*, 2024, 75(4): 1241-1255. (in Chinese)
- Liu, X., Zhang, W., Tong, X., et al. MolFilterGAN: A progressively augmented generative adversarial network for triaging AI-designed molecules. *Journal of Cheminformatics*, 2023, 15(1): 42.
- Lu, D., Wang, H., Chen, M., et al. 86 PFLOPS deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy. *Computer Physics Communications*, Amsterdam, 2021, 259: 107624.
- Lu, J., Pan, J., Mo, Y., et al. Automated intelligent platforms for high-throughput chemical synthesis. *Artificial Intelli-*

- gence Chemistry, 2024, 2(1): 100057.
- Lv, W., Zhou, Z., Zhang, Q., et al. Study on the mechanism of surfactant flooding: Effect of betaine structure. *Advances in Geo-Energy Research*, 2023, 146-158.
- Maia, K. C. B., Densy dos Santos Francisco, A., Moreira, M. P., et al. Advancements in surfactant carriers for enhanced oil recovery: Mechanisms, challenges, and opportunities. *ACS Omega*, 2024, 9(35): 36874-36903.
- Majewski, M., Pérez, A., Thölke, P., et al. Machine learning coarse-grained potentials of protein thermodynamics. *Nature Communications*, 2023, 14(1): 5739.
- Mao, Y., He, Q., Zhao, X. Designing complex architected materials with generative adversarial networks. *Science Advances*, 2020, 6(17): eaaz4169.
- Matsuzaka, Y., Uesawa, Y. A deep learning-based quantitative structure-activity relationship system construct prediction model of agonist and antagonist with high performance. *International Journal of Molecular Sciences*, 2022, 23(4): 2141.
- McLoughlin, K. S., Shi, D., Mast, J. E., et al. Generative molecular design and experimental validation of selective histamine H1 inhibitors. *bioRxiv*(p. 2023.02.14.528391), 2023.
- Millemann, R. E., Haynes, R. J., Boggs, T. A., et al. Enhanced oil recovery: Environmental issues and state regulatory programs. *Environment International*, 1982, 7(3): 165-177.
- Mouallem, J., Raza, A., Glatz, G., et al. Estimation of CO₂-Brine interfacial tension using machine learning: Implications for CO₂ geo-storage. *Journal of Molecular Liquids*, 2024, 393: 123672.
- Nguyen, T., Karolak, A. Transformer graph variational autoencoder for generative molecular design. *Biophysical Journal*, 2025.
- Nnadili, M., Okafor, A. N., Olayiwola, T., et al. Surfactant-specific AI-driven molecular design: Integrating generative models, predictive modeling, and Reinforcement Learning for tailored surfactant synthesis. *Industrial & Engineering Chemistry Research*, 2024, 63(14): 6313-6324.
- Nnadili, M., Okafor, A., Olayiwola, T., et al. Generative AI-driven molecular design: Combining predictive models and reinforcement learning for tailored molecule generation. *ChemRxiv*, 2023.
- O'Neill, S. AI-driven robotic laboratories show promise. *Engineering*, 2021, 7(10): 1351-1353.
- Peng, G. C. Y., Alber, M., Buganza Tepole, A., et al. Multiscale modeling meets machine learning: What can we learn? *Archives of Computational Methods in Engineering*, 2021, 28(3): 1017-1037.
- Peter, C., Kremer, K. Multiscale simulation of soft matter systems—from the atomistic to the coarse-grained level and back. *Soft Matter*, 2009, 5(22): 4357-4366.
- Qu, J., Wan, Y., Tian, M., et al. Microemulsions based on diverse surfactant molecular structure: Comparative analysis and mechanistic study. *Processes*, 2023, 11(12): 3409.
- Rahmanian, F., Flowers, J., Guevarra, D., et al. Enabling modular autonomous feedback-loops in materials science through hierarchical experimental laboratory automation and orchestration. *Advanced Materials Interfaces*, 2022, 9(8): 2101987.
- Rashidi-Khaniabadi, A., Rashidi-Khaniabadi, E., Amiri-Ramsheh, B., et al. Modeling interfacial tension of surfactant-hydrocarbon systems using robust tree-based machine learning algorithms. *SCIENTIFIC REPORTS*, Berlin, 2023, 13(1): 10836.
- Roch, L. M., Häse, F., Kreisbeck, C., et al. ChemOS: Orchestrating autonomous experimentation. *Science Robotics*, 2018, 3(19): eaat5559.
- Rosen, M. J., *Surfactants and Interfacial Phenomena*. Hoboken, USA, John Wiley & Sons, 2012.
- Saberi, H., Karimian, M., Esmailnezhad, E. Performance evaluation of ferro-fluids flooding in enhanced oil recovery operations based on machine learning. *Engineering Applications of Artificial Intelligence*, 2024, 132: 107908.
- Sabhapondit, A., Borthakur, A., Haque, I. Characterization of acrylamide polymers for enhanced oil recovery. *Journal of Applied Polymer Science*, 2003, 87(12): 1869-1878.
- Sadeghi, S., Canty, R. B., Mukhin, N., et al. Engineering a sustainable future: Harnessing automation, robotics, and artificial intelligence with self-driving laboratories, *ACS Sustainable Chemistry & Engineering*, 2024, 12(34): 12695-12707.
- Santo, K. P., Neimark, A. V. Dissipative particle dynamics simulations in colloid and interface science: A review. *Advances in Colloid and Interface Science*, 2021, 298: 102545.
- Schütt, K., Kindermans, P. J., Saucedo Felix, H. E., et al., SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *ArXiv Preprint ArXiv: 1706.08566*, 2017.
- Seifrid, M., Pollice, R., Aguilar-Granda, A., et al. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 2022, 55(17): 2454-2466.
- Shakeel, M., Pourafshary, P., Hashmet, M. R., et al. Application of machine learning techniques to predict viscosity of polymer solutions for enhanced oil recovery. *Energy Systems*, 2023.
- Shang, C., Kang, P., Liu, Z. Development and application of atomic simulation software based on machine learning potentials. *Journal of the Chinese Ceramic Society*, 2023, 51(2): 476. (in Chinese)
- Shi, J., Wu, Z., Deng, Q., et al. Synthesis of hydrophobically associating polymer: Temperature resistance and salt tolerance properties. *Polymer Bulletin*, 2022, 79(7): 4581-4591.
- Shilko, E. V., Dmitriev, A. I., Balokhonov, R. R., et al. Multiscale modeling and computer-aided design of advanced materials with hierarchical structure. *Physical Mesomechanics*, 2024, 27(5): 493-517.
- Shinkle, E., Pachalieva, A., Bahl, R., et al. Thermodynamic transferability in coarse-grained force fields using graph neural networks. *Journal of Chemical Theory and Com-*

- putation, 2024, 20(23): 10524-10539.
- Sipl, W., Robaa, D., Qsar/Qspr, in *Applied Chemoinformatics*, edited by Thomas, E., Johann, G., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp. 9-52, 2018.
- Sun, Q., Ertekin, T., Zhang, M., et al. A comprehensive techno-economic assessment of alkali-surfactant-polymer flooding processes using data-driven approaches. *Energy Reports*, 2021, 7: 2681-2702.
- Szymanski, N. J., Rendy, B., Fei, Y., et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*. 2023, 624(7990): 86-91.
- Tartakovskiy, A. M., Panchenko, A. Pairwise force smoothed particle hydrodynamics model for multiphase flow: Surface tension and contact line dynamics. *Journal of Computational Physics*, 2016, 305: 1119-1146.
- Taylor, K. C., Nasr-El-Din, H. A. Water-soluble hydrophobically associating polymers for improved oil recovery: A literature review. *Journal of Petroleum Science and Engineering*, 1998, 19(3): 265-280.
- Tom, G., Schmid, S. P., Baird, S. G., et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 2024, 124(16): 9633-9732.
- Tropsha, A., Isayev, O., Varnek, A., et al. Integrating QSAR modelling and deep learning in drug discovery: The emergence of deep QSAR. *Nature Reviews Drug Discovery*, 2024, 23(2): 141-155.
- Vamathevan, J., Clark, D., Czodrowski, P., et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 2019, 18(6): 463-477.
- Wan, W., Zhao, J., Harwell, J. H., et al. Characterization of crude oil equivalent alkane carbon number (EACN) for surfactant flooding design. *Journal of Dispersion Science and Technology*, 2016, 37(2): 280-287.
- Wang, H., Zhang, L., Han, J., et al. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 2018, 228: 178-184.
- Wang, J., Sun, H., Huang, Z., et al. Current status and prospect of percolation theory and development technologies of oil and gas reservoirs. *Science and Technology Foresight*, 2023, 2(2): 131-144. (in Chinese)
- Wang, L., Zhang, Y., Zou, R., et al., Applications of molecular dynamics simulation in studying shale oil reservoirs at the nanoscale: Advances, challenges and perspectives. *Petroleum Science*, 2025, 22(1): 234-254.
- Wang, S., Feng, Q., Javadpour, F., et al. Multiscale modeling of gas transport in shale matrix: An integrated study of molecular dynamics and rigid-pore-network model. *SPE Journal*, 2020, 25(3): 1416-1442.
- Wang, Z., Chen, A., Tao, K., et al., MatGPT: A vane of materials informatics from past, present, to future. *Advanced Materials*, 2024, 36(6): 2306733.
- Waqar, A., Othman, I., Shafiq, N., et al. Applications of AI in oil and gas projects towards sustainable development: A systematic literature review. *Artificial Intelligence Review*, 2023, 56(11): 12771-12798.
- Wen, T., Zhang, L., Wang, H., et al. Deep potentials for materials science. *Materials Futures*, 2022, 1(2): 022601.
- Xiong, B., Loss, R. D., Shields, D., et al. Polyacrylamide degradation and its implications in environmental systems. *NPJ Clean Water*, 2018, 1: 17.
- Xu, J., Cao, X., Hu, P. Accelerating metadynamics-based free-energy calculations with adaptive machine learning potentials. *Journal of Chemical Theory and Computation*, 2021, 17(7): 4465-4476.
- Yang, X., Wang, Y., Byrne, R., et al. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 2019, 119(18): 10520-10594.
- Yao, S., Song, J., Feng, Z., et al. Advances in deep learning-based 3D molecular generative models. *Scientia Sinica Chimica*, 2023, 53(2): 174-195. (in Chinese)
- Yousefmarzi, F., Haratian, A., Mahdavi Kalatehno, J., et al. Machine learning approaches for estimating interfacial tension between oil/gas and oil/water systems: A performance analysis. *Scientific Reports*, 2024, 14(1): 858.
- Yuan, S., Han, H., Wang, H., et al. Research progress and potential of new enhanced oil recovery methods in oilfield development. *Petroleum Exploration and Development*, 2024, 51(4): 963-980.
- Zeng, J., Zhang, D., Lu, D., et al. DeePMD-kit v2: A software package for deep potential models. *The Journal of Chemical Physics*, 2023, 159(5): 054801.
- Zerpa, L. E., Queipo, N. V., Pintos, S., et al. An optimization methodology of alkaline-surfactant-polymer flooding processes using field scale numerical simulation and multiple surrogates. *Journal of Petroleum Science and Engineering*, 2005, 47(3): 197-208.
- Zhang, L., Han, J., Wang, H., et al. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Physical Review Letters*, 2018a, 120(14): 143001.
- Zhang, L., Han, J., Wang, H., et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *ArXiv Preprint ArXiv: 1805.09003*, 2018b.
- Zhang, R., Nolte, D., Sanchez-Villalobos, C., et al. Topological regression as an interpretable and efficient tool for quantitative structure-activity relationship modeling. *Nature Communications*, 2024, 15(1): 5072.
- Zhou, J., Huang, M. Navigating the landscape of enzyme design: From molecular simulations to machine learning. *Chemical Society Reviews*, 2024, 53(16): 8202-8239.