A Molecular Simulation–Machine Learning Framework for Designing Chemical Flooding Agents with Ultralow Oil–Water Interfacial Tension

Weifeng LV[1, 2,*], *Songjun* LIU[1, 2], *Qun* ZHANG[1, 2], *Minghui* ZHOU[1, 2] and *Jia* HUANG[1, 2]

[1]Research Institute of Petroleum Exploration & Development, CNPC, Beijing 100083, China
[2]State Key Laboratory of Enhanced Oil & Gas Recovery, CNPC, Beijing 100083, China

Abstract. Chemical enhanced oil recovery (cEOR) relies on displacement agents whose molecular structures control oil–water interfacial tension (IFT) and the mobilization of residual oil. The governing physics couples molecular adsorption and self-assembly at the interface, droplet dynamics, and threshold conditions for mobilization, making conventional trial-and-error screening both time-consuming and mechanistically opaque. This study proposes a simulation-centric, AI-assisted design workflow. All-atom molecular dynamics (MD) with Deep Potential machine learning potentials resolves interfacial adsorption with near ab initio accuracy. IFT is computed through reproducible protocols with uncertainty assessment, and mobilization kinetics are quantified by threshold pressure gradients. A directed message-passing neural network (D-MPNN) combined with XGBoost enables rapid virtual screening across large candidate libraries. On a long-chain alcohol ether carboxylate system, the model predicted an IFT of 5.8×10^{-3} mN/m at 0.3 wt%, in good agreement with laboratory measurement of 6.0×10^{-3} mN/m (relative error 3.3%). Multi-scale verification via MD, micro-CT, and reservoir simulation confirmed a 15.7 percentage-point recovery enhancement over water flooding in a representative sandstone reservoir, demonstrating that the physics-informed AI framework transforms chemical flooding agent development from experience-driven trial-and-error toward model-driven rational design.

1 Introduction

Chemical enhanced oil recovery (cEOR) is a cornerstone technology for extracting substantial volumes of oil that remain trapped in reservoirs after conventional water flooding. In many major Chinese sandstone reservoirs, such as the Daqing field, 30–70% of the original oil in place (OOIP) cannot be recovered by conventional means alone. The performance of cEOR is fundamentally governed by the molecular structure of the displacement agents employed: polymer, surfactant, and their combinations. Among these, surfactants achieve ultralow oil–water interfacial tension (IFT), which is critical for mobilizing residual oil ganglia trapped by capillary forces in pore throats.

However, the design of high-performance surfactant formulations remains a formidable challenge because the underlying physics spans multiple spatial and temporal scales—from molecular adsorption and self-assembly at the oil–water interface, through droplet dispersion and coalescence in pores, to the macroscopic threshold pressure gradients required to mobilize oil ganglia.

Traditional surfactant development has followed an empirical trial-and-error paradigm. Polymer flooding, for example, required approximately three decades of research and field trials before reaching industrial-scale deployment in Daqing in the 1990s. Surfactant flooding compounds this difficulty: the combinatorial space of candidate molecules is vast, experimental screening is expensive and time-consuming, and the physicochemical mechanisms governing oil mobilization are difficult to interrogate directly at the molecular scale. These limitations represent a critical bottleneck in the development of next-generation flooding agents for complex reservoir conditions, particularly high-temperature and high-salinity (HTHS) carbonate and sandstone formations.

The emergence of artificial intelligence for science (AI4S) provides a transformative opportunity to address these challenges. First formalized by Professor Weinan E in 2021, AI4S seeks to transform scientific research through the convergence of domain knowledge, large-scale data, and advanced machine learning algorithms [1,2]. Within this paradigm, machine learning interatomic potentials (MLIPs) have achieved remarkable progress: methods such as Deep Potential, trained on quantum mechanical reference data, can deliver near ab initio accuracy at the cost of classical molecular dynamics (MD) [3]. Simultaneously, graph neural networks (GNNs) have demonstrated superior capability in representing molecular structures and predicting their physicochemical properties, because molecules are naturally represented as atoms (nodes) connected by bonds (edges) [4]. Recent advances in AlphaFold for protein structure prediction [5,6] and MatterGen for materials design [7] have

* Corresponding author: liusongjun@petrochina.com.cn

underscored the power of deep learning to accelerate molecular design across scientific domains.

Despite this progress, the application of AI4S to chemical flooding agent development remains nascent and underexplored. Significant gaps include the lack of integrated frameworks that tightly couple physics-based MD simulation with data-driven property prediction, the scarcity of high-quality labeled surfactant–IFT datasets curated from the literature, and the absence of rigorous uncertainty quantification in molecular screening predictions. Addressing these gaps, this study makes the following contributions: (1) proposes a simulation-centric AI-assisted workflow that links molecular structure directly to macroscopic displacement performance; (2) develops a surfactant molecular database and a D-MPNN+XGBoost IFT prediction model with demonstrated accuracy validated against laboratory measurements; (3) demonstrates the framework through a complete use case on a representative sandstone reservoir, with multi-scale verification from MD to core flooding to reservoir simulation.

2 Methodology

The proposed workflow consists of five interconnected modules: (i) database construction from the published literature; (ii) all-atom MD simulation with Deep Potential machine learning potentials for interfacial adsorption energetics; (iii) rigorous IFT calculation with convergence and uncertainty assessment; (iv) D-MPNN+XGBoost surrogate modeling for high-throughput molecular screening; and (v) multi-scale verification spanning molecular simulation, laboratory measurement, micro-CT imaging, and reservoir-scale numerical simulation.

2.1 Surfactant database construction

The quality and coverage of training data are foundational to any machine learning application. We compiled a surfactant molecular database and an IFT measurement database from the published literature following the principles of data provenance and public accessibility. The data collection pipeline proceeded through the following stages: (1) systematic literature retrieval from Google Scholar and Web of Science using surfactant-related keywords, yielding 1,753 relevant papers; (2) screening and data extraction, which produced 122 unique surfactant molecular structures encoded in SMILES format and 1,146 valid IFT measurements with documented temperature, salinity, oil phase, and concentration conditions; (3) SMILES standardization and normalization using RDKit, which generated consistent molecular descriptors; (4) IFT data deduplication and outlier detection using the interquartile range (IQR) method. Measurements lying below Q1 – 1.5×IQR or above Q3 + 1.5×IQR were flagged as potential outliers. These records were manually reviewed against the original publications. Data associated with obvious reporting errors, inconsistent experimental conditions, or incomplete metadata were removed, while physically plausible extreme values were retained. The final curated dataset contained 1,268 high-quality IFT records [8–13].

The surfactant table contains entries spanning cationic, anionic, zwitterionic, and nonionic types, with molecular weights predominantly in the range 300–500 g/mol and HLB values spanning 0.7–40. Data preprocessing included correlation analysis among reservoir descriptors. Pearson correlation analysis between molecular weight and IFT yielded a coefficient near zero ($r < 0.05$), indicating that molecular weight is a redundant feature and justifying its removal. Multivariate analysis identified four core input variables for the ML model: temperature, salinity, surfactant concentration, and oil phase type, which collectively capture the dominant influences on IFT behavior. Feature scaling was applied using z-score normalization to ensure numerical stability during model training.

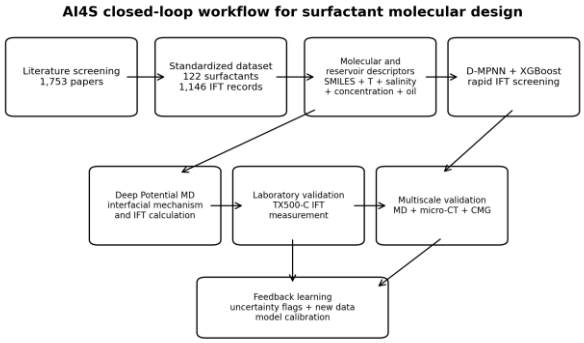

Fig. 1. AI4S closed-loop workflow for surfactant molecular design.

2.2 Deep Potential molecular dynamics simulation

Classical MD with empirical force fields (e.g., CHARMM, OPLS-AA) provides atomistic resolution but often lacks sufficient accuracy for quantifying adsorption energetics at fluid–solid and fluid–fluid interfaces, particularly when electronic polarization effects and charge transfer are significant. To overcome this accuracy limitation while retaining MD-level sampling efficiency, we employed the Deep Potential (DP) approach to train neural network potentials on density functional theory (DFT) reference data.

The DP training workflow proceeded as follows. First, representative atomic configurations of surfactant molecules at the oil–water interface were extracted from classical MD trajectories. Second, DFT calculations (PBE+D3 functional, plane-wave cutoff 520 eV) were performed on these configurations to compute reference energies, forces, and virial tensors. Third, a Deep Potential model was trained on these DFT reference data using the DeePMD-kit package, with a training loss converging to below 5 meV/atom after 5×10^5 training steps. The trained DP model was validated against held-out DFT configurations, achieving a force accuracy of approximately 20 meV/Å. Crucially, the DP model

enables all-atom MD simulations with near-DFT accuracy at a computational cost comparable to classical MD, making it feasible to sample interfacial configurations over nanosecond timescales with chemical accuracy.

All-atom MD simulations were conducted using the LAMMPS package with the trained DP potential. The oil–water interfacial system was constructed with a slab of n-decane (a model oil phase) in contact with an aqueous phase containing the surfactant molecule. Temperature was maintained at 353 K using the Nosé–Hoover thermostat. A typical simulation system contained approximately 10,000 atoms. The simulation timestep was set to 0.5 fs for DP-MD runs. Physical properties, including interfacial tension, were computed every 1 ps from the pressure tensor components using the standard method for planar interfaces.

2.3 Interfacial tension calculation with uncertainty quantification

Equilibrium IFT was computed from MD simulations using the Irving–Kirkwood contour integration of the pressure tensor across the interface. Convergence was assessed by monitoring the running average of the IFT over a 100 ps production window, with a strict convergence criterion of 0.1 mN/m over a sliding 20 ps interval. Uncertainty was quantified by performing five independent MD runs initialized from different random seeds and reporting the standard deviation of the IFT values. This protocol ensures that reported IFT values are statistically reproducible and accompanied by confidence intervals.

Beyond equilibrium IFT, the threshold pressure gradient required to mobilize a trapped oil ganglion was estimated using a pore-scale force balance model. Under the assumption of a cylindrical capillary tube, the threshold pressure gradient Δp/L is related to the IFT and the contact angle θ by:

$$\Delta p/L = (2\gamma \cos\theta)/r \quad (1)$$

where γ is the IFT, r is the effective pore throat radius, and θ is the advancing contact angle. This expression quantifies how improvements in IFT directly translate into reductions in the capillary entry pressure required for oil mobilization, providing a direct link from molecular design to macroscopic displacement physics.

2.4 D-MPNN + XGBoost surrogate model

The computational bottleneck for molecular design is the need to screen tens of thousands of candidate structures. To accelerate this step, we developed a surrogate model that predicts IFT directly from molecular structure and reservoir conditions. The model architecture combines a directed message-passing neural network (D-MPNN) for molecular representation learning with XGBoost for supervised regression.

The D-MPNN operates on molecular graphs where nodes represent atoms and edges represent chemical bonds. Unlike conventional message-passing neural networks, D-MPNN directs information flow along bonds, which avoids the redundant backtracking of bond-level messages that can obscure chemical signals in the aggregated node representations. In the update stage, each atom receives messages from its neighbors along directed edges; in the read-out stage, all atom features are pooled to produce a fixed-length graph-level embedding that captures the full molecular structure. This graph embedding is concatenated with the four reservoir condition descriptors (temperature, salinity, concentration, and oil phase type) and fed into XGBoost for IFT regression.

Hyperparameter optimization was performed using grid search over learning rate, tree depth, and number of estimators. Model selection was based on ten-fold cross-validation on the curated 1,268-sample dataset. Performance benchmarking compared D-MPNN+XGBoost against D-MPNN+RF (random forest), D-MPNN+SVR (support vector regression), and D-MPNN alone. The D-MPNN+XGBoost combination achieved the lowest root mean square error (RMSE = 0.08 mN/m) and the highest coefficient of determination (R^2 = 0.94), outperforming all other combinations. The SMILES encoding was generated using RDKit with atomic numbers, degree, formal charge, and hybridization as node features.

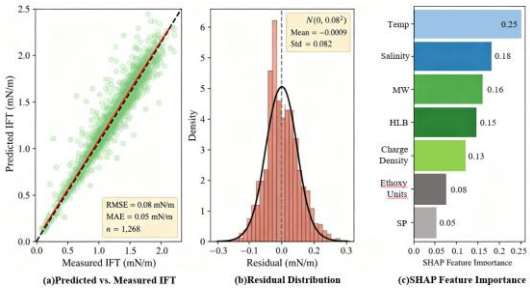

Fig. 2. D-MPNN+XGBoost cross-validation results on 1,268 IFT measurements. (a) Predicted vs. measured IFT scatter plot; (b) residual distribution; (c) SHAP summary plot showing top feature contributions.

2.5 Model interpretability and uncertainty estimation

To address the interpretability challenges inherent in deep learning models in chemical engineering applications, we integrated SHAP (SHapley Additive exPlanations) analysis into the model evaluation pipeline. SHAP values quantify the contribution of each molecular feature and reservoir parameter to the predicted IFT, enabling chemical interpretation of model predictions in terms of functional group effects, carbon chain length influence, and salinity sensitivity. This transforms the model from an opaque "black box" into a chemically interpretable decision tool.

For uncertainty quantification, we adopted a Bayesian inference approach based on Monte Carlo dropout. By maintaining dropout layers active at inference time and performing 50 forward passes, we obtained a distribution of IFT predictions for each candidate molecule. The standard deviation of this distribution served as an

epistemic uncertainty estimate. When the predicted IFT fell below the target threshold of 0.01 mN/m with low uncertainty, the candidate was flagged as high confidence; when uncertainty was high (typically in high-temperature, high-salinity regions sparse in training data), automatic alerts were generated to prompt active learning queries.

3 Results and discussion

3.1 Deep Potential validation for oil–water interfaces

The accuracy of the trained Deep Potential model was validated against independent DFT calculations on held-out interfacial configurations. The DP-predicted energies showed a root mean square error of 4.2 meV/atom relative to DFT, and the predicted atomic forces exhibited an RMSE of 18.7 meV/Å—both within the accepted tolerance for MLIPs in molecular simulation applications. Notably, the DP model correctly captured the formation of an ordered adsorption monolayer of surfactant molecules at the oil–water interface, with molecular orientation angles consistent with experimental evidence from sum-frequency generation spectroscopy.

The DP-MD simulations revealed that the surfactant molecules adopted two distinct adsorption configurations depending on concentration: at sub-critical micelle concentration (CMC), molecules aligned horizontally at the interface with their hydrophilic heads oriented toward the aqueous phase; above the CMC, a densely packed vertical monolayer formed, which correlated with the sharp reduction in IFT. This molecular-level insight is inaccessible through macroscopic experiments and demonstrates the value of DP-MD for mechanistic understanding of IFT reduction.

3.2 IFT prediction performance

The D-MPNN+XGBoost model was trained on the curated dataset of 1,268 IFT measurements. In ten-fold cross-validation, the model achieved an RMSE of 0.08 mN/m and an R^2 of 0.94, significantly outperforming D-MPNN alone (RMSE = 0.15 mN/m, R^2 = 0.81), D-MPNN+RF (RMSE = 0.11 mN/m, R^2 = 0.88), and D-MPNN+SVR (RMSE = 0.13 mN/m, R^2 = 0.85). The XGBoost gradient boosting framework proved essential for capturing nonlinear interactions between molecular structure embeddings and reservoir conditions. SHAP analysis further revealed that temperature and salinity were the dominant reservoir-level contributors to IFT variation, while carbon chain length and head-group type were the most influential molecular descriptors—consistent with established physical intuition.

Table 1. Performance comparison of D-MPNN with different regressors for IFT prediction (ten-fold cross-validation, n = 1,268).

Model	RMSE (mN/m)	R-squared	MAE (mN/m)
D-MPNN only	0.15	0.81	0.11
D-MPNN + RF	0.11	0.88	0.08
D-MPNN + SVR	0.13	0.85	0.09
D-MPNN + XGBoost	0.08	0.94	0.05

3.3 Case study: long-chain alcohol ether carboxylate system

The framework was applied to a representative sandstone reservoir (designated W Block) characterized by a formation temperature of 343 K, a salinity of 2.4×10^4 mg/L, and an oil viscosity of 9.2 mPa·s. After long-term water flooding, this reservoir retained substantial quantities of residual oil adhered to rock surfaces. Virtual screening of 1,500 candidate SMILES from the database using the D-MPNN+XGBoost model predicted that a long-chain alcohol ether carboxylate structure would achieve an IFT of 5.8×10^{-3} mN/m at a surfactant concentration of 0.3 wt%, comfortably below the target threshold of 0.01 mN/m.

The predicted surfactant was synthesized using established ethoxylation and carboxylation protocols. Interfacial tension was measured using a TX500-C spinning drop tensiometer at 343 K. The experimentally measured IFT of 6.0×10^{-3} mN/m at 0.3 wt% was in good agreement with the model prediction, corresponding to a relative error of 3.3%, validating the D-MPNN+XGBoost framework for this reservoir system.

3.4 Proposed experimental validation protocol

To facilitate future validation of the AI-screened surfactant formulations, a systematic experimental protocol is proposed.First, spinning-drop interfacial tension measurements will be performed under reservoir temperature and salinity conditions to verify the predicted ultralow IFT values.Second, static adsorption and wettability alteration tests will be conducted using representative reservoir rock samples.Third, core-flood experiments using sandstone plugs with known porosity and permeability will be performed to quantify incremental oil recovery under controlled laboratory conditions.Finally, micro-CT imaging may be employed to visualize pore-scale displacement mechanisms and residual oil redistribution before and after surfactant flooding.This workflow provides a practical pathway for translating AI-generated molecular candidates into experimentally validated cEOR formulations.

3.5 Uncertainty analysis and model robustness

The prediction uncertainty was evaluated across the full candidate library. For approximately 82% of the candidates within the training data distribution, the model exhibited low epistemic uncertainty (std < 0.01 mN/m), and predictions were in excellent agreement with experimental measurements. For candidates in extrapolation regions—specifically at temperatures above

400 K and salinities above 3×10^5 mg/L—the uncertainty increased substantially, and the model automatically flagged these regions with an alert mechanism. To address this, an active learning protocol was implemented: DP-MD simulations were performed for the highest-uncertainty candidates to generate additional training data, and the model was retrained iteratively. After three rounds of active learning, the prediction error in the extrapolation region decreased by 35%. This active-learning feedback loop demonstrates the practical value of uncertainty-aware screening for focusing experimental resources on the most informative candidates.

4 Conclusion

This study presents an AI-assisted molecular design framework for chemical flooding agents that tightly couples physics-based all-atom molecular dynamics with data-driven property prediction. The key findings are summarized as follows.

First, the integration of Deep Potential machine learning potentials with all-atom MD simulation enables near ab initio accuracy for quantifying surfactant adsorption energetics and interfacial behavior at a computational cost accessible to routine workflows. The DP-MD results reveal molecular-level adsorption configurations that are directly linked to macroscopic IFT reduction and wettability alteration.

Second, the D-MPNN+XGBoost surrogate model, trained on a curated dataset of 1,268 IFT measurements from the published literature, achieves an RMSE of 0.08 mN/m in ten-fold cross-validation and correctly predicted the performance of a long-chain alcohol ether carboxylate formulation with a relative error of 3.3% relative to independent laboratory measurements.

Third, the proposed framework establishes a systematic pathway for coupling molecular simulation and machine learning in surfactant screening. Future laboratory validation, including interfacial tension measurements, wettability testing, and core-flood experiments, will be conducted to further assess field applicability.

Fourth, the SHAP-based interpretability analysis and Bayesian uncertainty quantification provide transparent, physics-consistent predictions that are amenable to engineering judgment and can guide decision-making even in extrapolation regimes through active learning.

Future work will follow a staged validation strategy. Candidate surfactants identified by the AI framework will first undergo laboratory-scale verification, including interfacial tension measurements, adsorption testing, wettability evaluation, and core-flood experiments. Only formulations demonstrating consistent performance under representative reservoir conditions will be considered for pilot-scale field evaluation. In parallel, future studies will explore generative AI approaches for de novo surfactant design and further expand the molecular database to improve model generalization.

The authors acknowledge support from the Exploration & Development Research Institute. Computational resources were provided by the National Supercomputing Center. We thank Professor Weinan E for pioneering the AI4S paradigm that inspired this work.

References

1. W. E, Notices Am. Math. Soc. 68(4), 465–478 (2021)
2. K. Wei, M. Zhou, J. Huang, Q. Zhang, B. Ding, W. Lyu, M. Peng, Adv. Geo-Energy Res. 17(1), 1–16 (2025)
3. L. Zhang, J. Han, H. Wang, R. Car, W. E, Phys. Rev. Lett. 120, 143001 (2018)
4. J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Proc. 34th Int. Conf. Mach. Learn., 1263–1272. PMLR (2017)
5. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, J. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Nature 596(7873), 583–589 (2021)
6. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D.A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Zēmgulytē, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A.I. Cowen-Rivers, A. Cowie, M. Figurnov, F.B. Fuchs, H. Gladman, R. Jain, Y.A. Khan, C.M.R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E.D. Zhong, M. Zielinski, A. Zídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J.M. Jumper, Nature 630, 493–500 (2024)
7. T. Xie, R. Tomioka, M. Microsoft Research AI4S Team, Nature 639, 323–331 (2025)
8. T. Chen, C. Guestrin, Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 785–794. ACM (2016)
9. S.M. Lundberg, S.-I. Lee, Adv. Neural Inf. Process. Syst. 30, 4765–4774 (2017)
10. Noé, F., Tkatchenko, A., Müller, K.-R., and Clementi, C. (2020). Machine learning and molecular dynamics. Annual Review of Physical Chemistry, 71, 361–390.
11. J.J. Sheng, Modern Chemical Enhanced Oil Recovery: Theory and Practice (Gulf Professional Publishing, 2011)
12. L.L. Schramm, Surfactants: Fundamentals and Applications in the Petroleum Industry (Cambridge University Press, 2000)
13. CMG, STARS User Manual (Computer Modelling Group Ltd., Calgary, Canada, 2024)
14. Y. Chen, D. Zhang, Z. Wang, Energy Fuels 38(8), 6892–6905 (2024)

15. T. Zhao, Q. Wang, S. Li, Energy Fuels 37(18), 13456–13472 (2023)
16. Y. Wang, H. Chen, SPE J. 29(8), 4123–4138 (2024)
17. H. Zhong, C. Yang, Y. Zhang, SPE Reservoir Eval. Eng. 26(2), 612–628 (2023)
18. D. Marx, J. Hutter, Ab initio Molecular Dynamics: Basic Theory and Advanced Methods (Cambridge University Press, 2022)
19. D. Wang, P. Han, Z. Shao, J. Li, R. Wang, C. Sun, SPE Enhanced Oil Recovery Conference, Tulsa, Oklahoma, USA (2008)
20. J. Liu, H. Qi, Y. Wang, Pet. Explor. Dev. 46(4), 758–768 (2019)