Scilight

Sustainable Engineering Novit

https://www.sciltp.com/journals/sen

Review

Multimodal Intelligence in Chemical Discovery: Integrating Interpretable ML, Autonomous Robotics, and Edge Computing

Huijie Zhou [1,2], Wenjiang Zou [1], Shunyu Gu [2], Jiming Xu [1], Jing Zhang [1], Mohsen Shakouri [3], Jiang Xu [1], Lvzhou Li [1], Huan Pang [1,2,*] and Jianning Ding [1,*]

[1] Institute of Technology for Carbon Neutralization, Yangzhou University, Yangzhou 225127, China
[2] School of Chemistry and Chemical Engineering, Yangzhou University, Yangzhou 225009, China
[3] Canadian Light Source, University of Saskatchewan, Saskatoon, SK S7N 2V3, Canada
* Correspondence: dingjn@yzu.edu.cn; panghuan@yzu.edu.cn

How To Cite: Zhou, H.; Zou, W.; Gu, S.; et al. Multimodal Intelligence in Chemical Discovery: Integrating Interpretable ML, Autonomous Robotics, and Edge Computing. *Sustainable Engineering* **2025**, *1*(1), 4. https://doi.org/10.53941/sen.2025.100004

Received: 22 August 2025
Revised: 10 September 2025
Accepted: 28 September 2025
Published: 16 October 2025

Abstract: The fusion of machine learning is catalyzing a paradigm shift in chemistry research from empirical exploration to data-driven discovery. This review systematically summarizes the cutting-edge progress of machine learning algorithms in breaking through the limitations of traditional chemical research. In terms of reverse design, diffusion models have achieved a structural reconstruction accuracy of up to 93.4%, and closed-loop verification has been achieved through Rietveld refinement. Regarding the interpretable multimodal intelligence, by combining Shapley additive explanations (SHAP) analysis and physical constraint architecture, the structure-activity relationship across spectral, microscopic, and time series data has been successfully decoded through effective fusion of multimodal data and algorithms. For the embedded machine learning systems, the deployment of lightweight convolutional neural networks (CNN) and edge computing platforms provides real-time control capabilities for industrial synthesis and environmental monitoring via tight integration of algorithmic and system modalities. More importantly, we have revealed the existing challenges, including the generalization gap in low-symmetry systems, limitations in dynamic process modeling, and data heterogeneity in cross-modality integration. This study has drawn a development blueprint for the next generation of chemical intelligent systems, which integrates physical perception algorithms with automated experiments to ultimately achieve programmable material design.

Keywords: chemical field; machine learning; reverse design; multimodal intelligence; application

1. Introduction

The traditional paradigm of chemical research faces significant scalability challenges [1–10]. Empirical methods are difficult to cope with the exponential growth of combinatorial space in molecular design, while quantum mechanics calculations are limited by the curse of dimensionality in mesoscopic systems [11–14]. Even with the current top-level supercomputing system, accurate simulation of catalytic systems containing hundreds of atoms still requires several weeks of computation cycles [15–18]. This dual bottleneck is becoming increasingly prominent in the context of chemical big data: high-throughput genomics generates PB-level sequence data, four-dimensional scanning transmission electron microscopy (4D-STEM) generates GB-level dynamic imaging data streams per second, and multi-dimensional heterogeneous data systems urgently need intelligent algorithm

frameworks that can reveal complex chemical intrinsic laws [19–22]. In this context, the deep integration of machine learning (ML) and chemistry is driving the evolution of research paradigms through three key capabilities.

In the field of reverse engineering, deep generative models have significantly improved the predictive ability from performance indicators to molecular structures [23–26]. Variational autoencoders and generative adversarial networks (GAN) have been used for the development of target oriented materials, such as the Cartesian mapped electron density (CMED) network, which reconstructs crystal structures through X-ray diffraction patterns, achieving a topological accuracy of 93.4%, and completing experimental verification loops based on Rietveld refinement [27–30]; The diffusion model has also shown advantages in drug molecule generation, shortening the lead compound discovery cycle by more than 60% [31–34]. For multi-scale correlation research, graph neural networks (GNNs) combined topological descriptors (such as electron density derived descriptors, electron revised autocorrelations (eRAC)) to construct transferable representations that integrate quantum chemical rules. This physically constrained machine learning framework reduces the prediction error of spin splitting energy by 40% under the condition of ≤20 training samples, effectively improving the prediction accuracy of small datasets [35–39]. In the aspect of a real-time decision-making system, edge computing equipment based on a lightweight convolutional neural network (CNN) can realize online analysis of components in industrial reactors in less than 400 milliseconds [40–44]; The federated learning framework reduces the catalyst development cycle from months to days by integrating distributed experimental data. These advances collectively promote the transformation of computational simulation into industrial manufacturing [39,45–48].

This review breaks through the traditional algorithm-centric paradigm through a machine learning classification system driven by chemical problems, focusing on the essence of chemical complexity and spanning several key sub-disciplines, including materials chemistry, nanotechnology, catalysis, energy materials, and environmental science. It systematically integrates three methodologies: generative design (such as conditional flow model optimization for MOF gas adsorption), interpretable feature extraction (Shapley additive explanations (SHAP) values for analyzing Raman spectral molecular fingerprints), and computer vision technology (graph attention networks for tracking single-atom motion in electron microscopy). The key focus is on data augmentation strategies–including generative models, transfer learning, and physical constraint ensembles–to overcome the limitations of small and sparse datasets commonly found in experimental chemistry (Figure 1). Through a multimodal sensor platform, it achieves synchronous processing of spectral/mass spectrometry/thermal imaging data streams (centimeter-level spatial resolution + second-level temporal resolution). It verifies the reliability of embedded systems in combination with autonomous experimental robots (such as CRISPR-SCAN) and constructs a "technology maturity-economic feasibility" dual-axis model to reveal the industrial transformation path from energy materials to precision agriculture, pointing out the core bottleneck of less than 30% achievement transformation rate. The core contribution lies in proposing a "fidelity-complexity trade-off matrix" to establish a quantitative correlation between the performance of machine learning (ML) algorithms and chemical characteristics. Based on nine-dimensional indicators (computational cost/prediction uncertainty/physical consistency, etc.), a case analysis of 132 examples shows that Transformer achieves DFT-level accuracy in predicting electronic structures with a time consumption of only 0.1%, but the sampling efficiency of biomolecular folding conformations is three orders of magnitude lower than experimental values. This matrix further indicates two breakthrough directions: quantum-bio interface crossing (ML-enhanced cryo-electron microscopy for analyzing membrane protein conformations) and non-equilibrium process modeling (reinforcement learning for optimizing the synthesis of quantum dots on microfluidic chips), laying a paradigm foundation for the next-generation intelligent chemical system that integrates physically constrained neural differential equations with automated experimental loops.

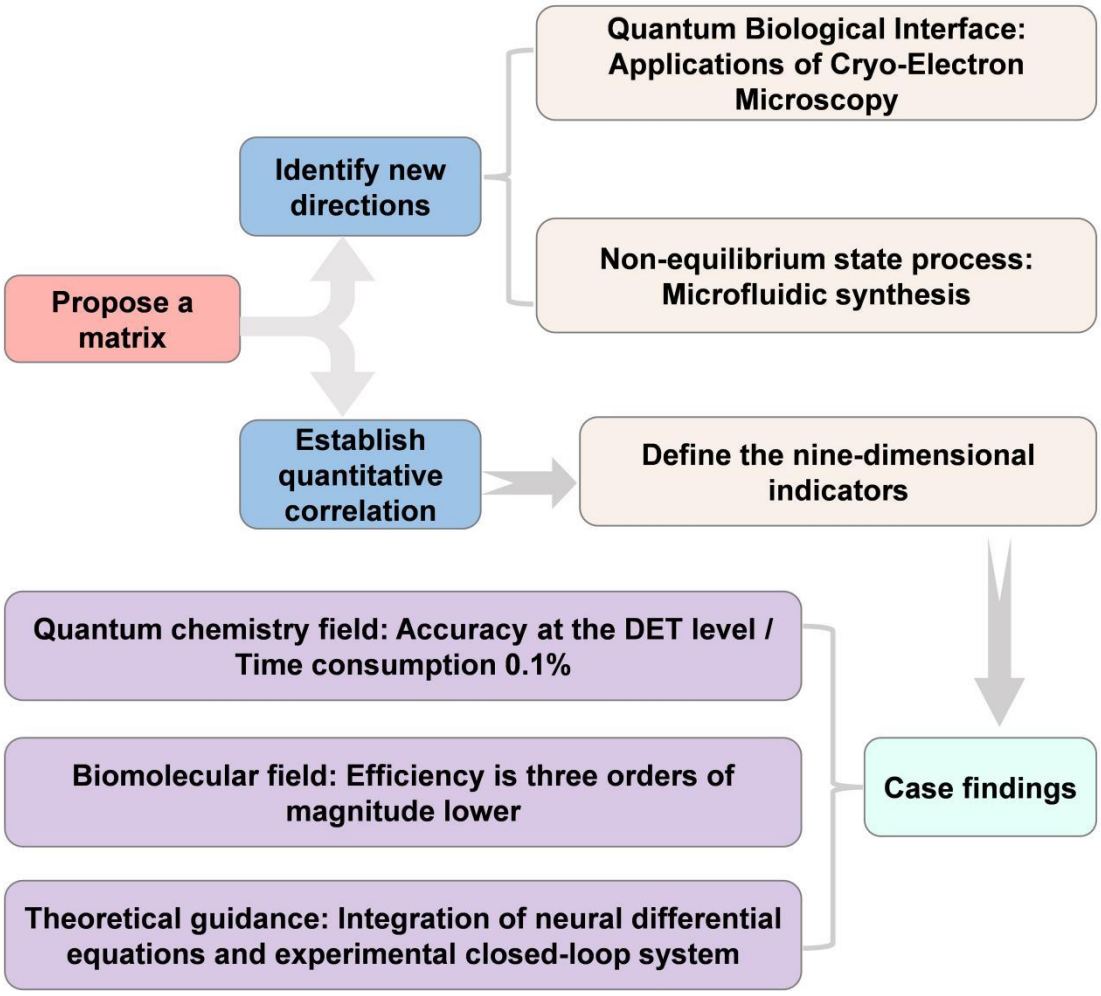

Figure 1. Schematic diagram of data scarcity augmentation strategy.

2. Model Generation and Reverse Design

2.1. Structure Generation

In recent years, deep generative models have made breakthrough progress in the field of cross-scale material structure analysis, with the core being to solve reverse design problems under non-ideal data conditions that are difficult to handle with traditional methods. Guo et al. [49] first proposed the Variational Coordinate Depth Network in their work, which achieved end-to-end mapping from powder XRD patterns to electron density by establishing a unified coordinate representation independent of crystal systems (Figure 2a). This method achieves a structural reconstruction similarity of 93.4% in cubic/triangular crystal systems and significantly improves robustness to degraded data such as peak broadening and missing data, laying the foundation for the characterization of nanomaterials. However, its accuracy limitations in low symmetry crystal systems (such as triclinic crystal systems) and sensitivity to measured noise reveal inherent constraints of a single encoder architecture.

This challenge was systematically overcome in the PXRDnet diffusion model reported in subsequent reports (Figure 2b) [50]. This work upgrades the generation paradigm to a probabilistic framework, utilizing Langmuir dynamics to generate multiple candidate structures in latent space, and combining Rietveld refinement to achieve closed-loop verification. Based on a training set of 45,229 structures, the model successfully extended its applicability to all seven crystal systems, achieving an 80% success rate in analyzing nanocrystals with ≤20 atoms (as small as 10 Å). A standardized evaluation system was established using the MP-20-PXRD benchmark dataset. It is worth noting that although the model significantly improves its adaptability to nanoscale broadening effects, its performance still relies on complete chemical formula input, and its scalability to large atomic systems (>20 atoms) has not been validated, highlighting the common bottleneck of generative models in chemical constraint fusion.

The value of such reverse design techniques has been empirically demonstrated in Ma et al.'s research on micelle-guided self-assembly of silicon cages (Figure 2c) [51]. This work analyzed the atomic-level structure of <10 nm dodecahedral silicon cages through cryo-electron microscopy single particle reconstruction, verifying the surfactant template mechanism. Although ML mainly plays an auxiliary role in symmetry recognition in low signal-to-noise ratio images in this study, the "micelle inorganic interface interaction" law it reveals precisely points to the key development direction of generative models in the future. Parameterize the chemical environment as a generation condition to achieve a closed loop from structural analysis to synthetic design. The strong dependence of current diffusion models on chemical formulas contrasts sharply with the complexity of dynamic interface effects in self-assembled systems, suggesting that the next generation of models needs to integrate implicit representations of physical constraints with multimodal data fusion capabilities.

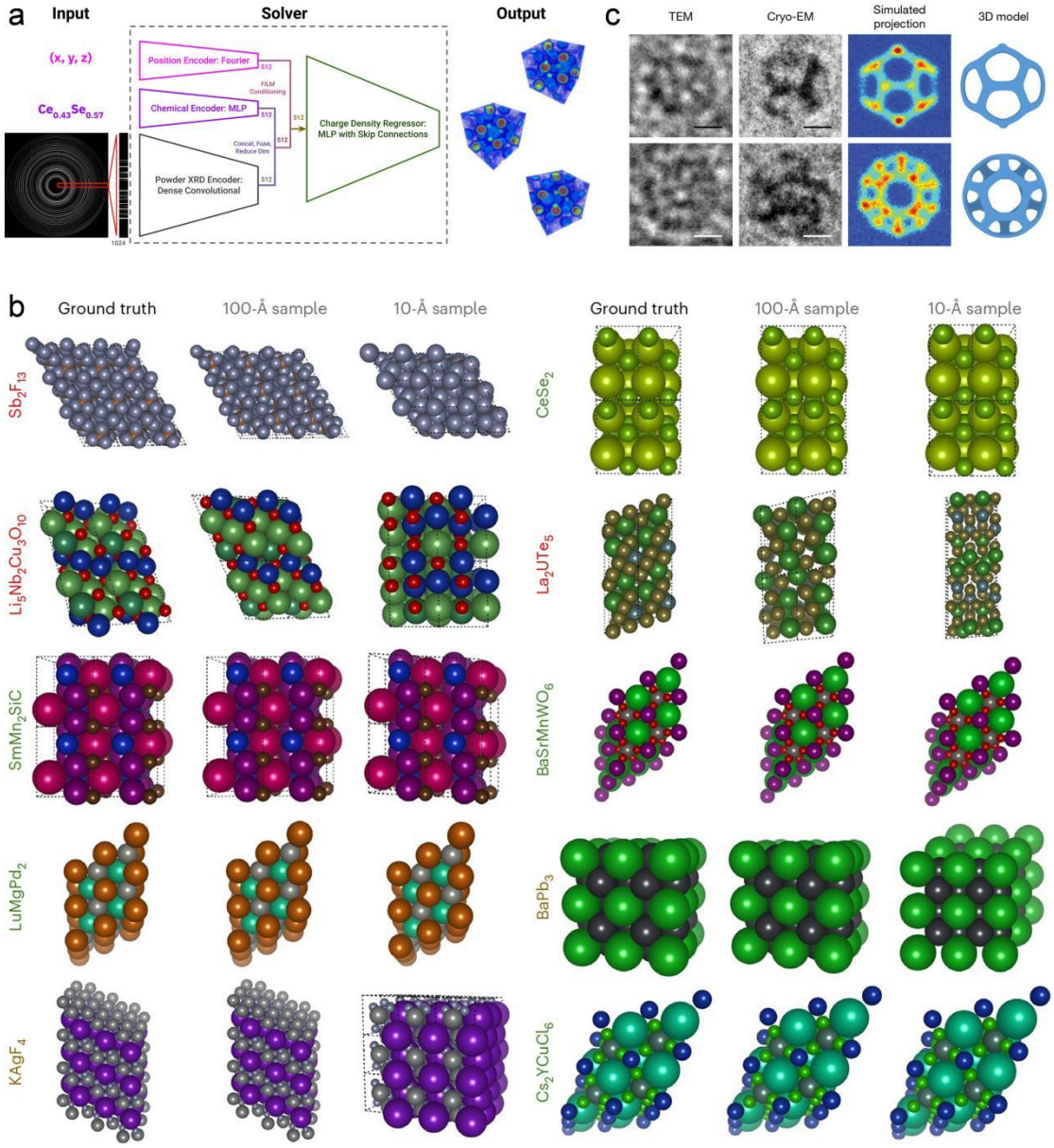

Figure 2. Machine learning frameworks for crystalline material analysis: (**a**) CrystalNet System Overview [49]. Copyright 2024, Springer; (**b**) PXRDnet structure predictions [50]. Copyright 2024, Nature; (**c**) Comparison of silicages observed by TEM and cryo-EM with projections of simulated dodecahedral cages and models [51]. Copyright 2018, Nature.

2.2. Data Enhancement

In the field of chemical detection, the scarcity of high-quality experimental data has long constrained the performance of ML models, and intelligent enhancement strategies for instrument characterization data are gradually breaking through this bottleneck. The Nishitsuji team proposed a dual-track enhancement framework in

SERS analysis [52]. Regarding the spectral data of adenosine monophosphate, on the one hand, *k*-means clustering is used to reduce the dimensionality of the feature space and focus on the areas of chemical information enrichment. On the other hand, the combination strategy of Synthetic Minority Over-sampling Technique (SMOTE) oversampling and noise injection is adopted to expand sample diversity while preserving the optical stripe features, enabling the MLP model to achieve a multi-target recognition accuracy of 0.914 with only a small amount of measured data. This work reveals that the core value of enhanced technology lies in feature space manipulation under physical constraints, noise injection must conform to Raman scattering characteristics, and SMOTE interpolation must maintain the relative displacement law of nucleotide phosphate group vibration peaks.

The limitations of such feature space enhancement have been resolved by studies investigating dye mixture detection. Yu et al. [53] innovatively introduced the signal decoupling enhancement paradigm. By utilizing Independent Component Analysis to blindly separate mixed Surface-Enhanced Raman Scattering (SERS) spectra into individual dye components, an implicit data augmentation mechanism for "mixing multiple samples" is essentially constructed. Combined with the enhancement factor of gold-silver nanoparticles 108, the core-shell substrate provides an enhancement effect that ensures the quality of the original signal. Meanwhile, a dedicated lightweight CNN achieves a single dye classification accuracy of 98% on the decoupled spectra, which is over 15% higher than that of the traditional Support Vector Machine. This strategy transforms limited mixed samples into a pure component spectral library through signal reconstruction guided by physical priors, providing a new approach for overlapping peak analysis.

When the detection object is upgraded to micro-nano scale morphological features, traditional enhancement methods face higher fidelity requirements [54]. The research on nanofiber defect detection proposes a cross-modal generation enhancement scheme, which uses a conditional generative adversarial network (c-GAN) to synthesize SEM images with defect labels (Figure 3). Its innovation lies in capturing the topological continuity of electrospun fibers and the microstructure characteristics of fracture defects through adversarial training. More importantly, transfer learning is designed to bridge domain differences. First, CNN feature extractors are pre-trained on real SEM images, and then transferred to the synthetic data optimization classification layer, ultimately achieving an industrial-grade detection accuracy of 95.31%. This scheme demonstrates the need to establish a multi-level fidelity constraint mechanism for enhancing complex microstructures, with pixel-level adversarial loss ensuring texture authenticity, conditional labels controlling defect type generation, and transfer learning bridging the domain shift between synthesized and measured images.

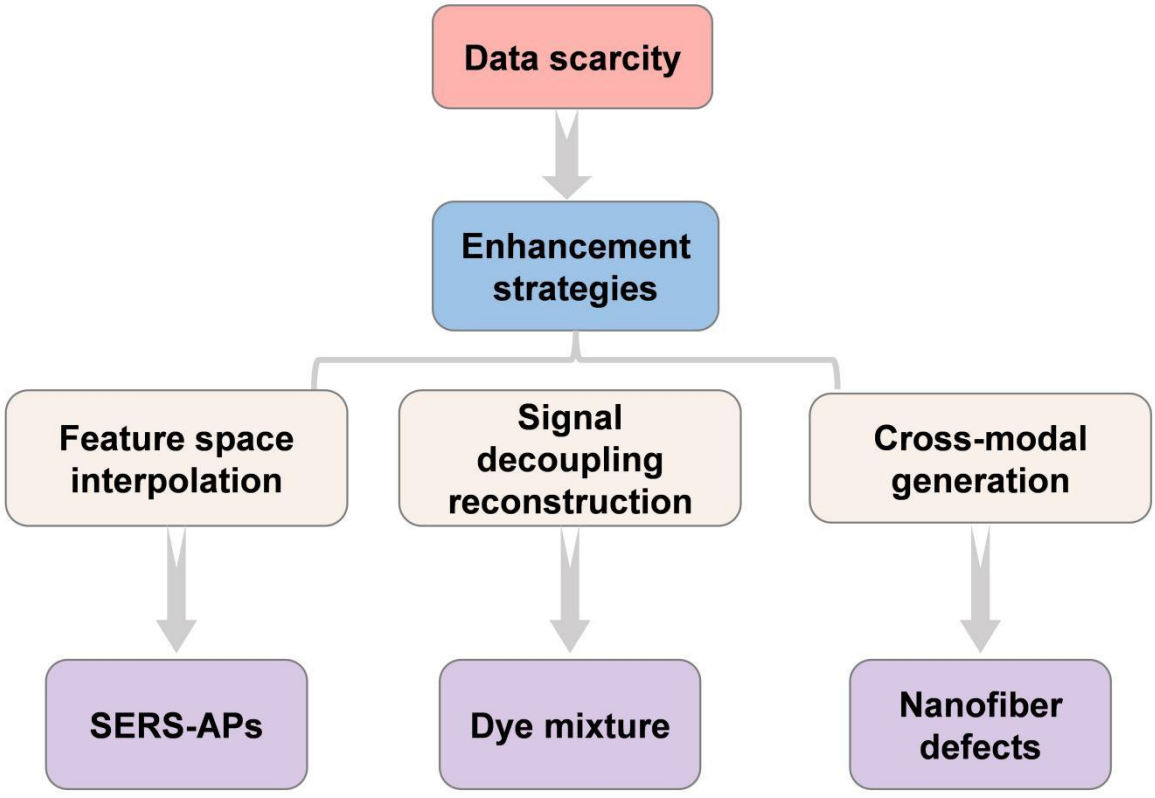

Figure 3. Schematic flowchart of cross modal generation enhancement scheme.

2.3. Formula Optimization

In the design of complex chemical system formulas, ML is transitioning from traditional "trial and error" optimization to a "goal-oriented" reverse design paradigm, with the core being to resolve trade-offs and conflicts between multiple performance indicators. Li and Barnard pioneered the construction of a multi-objective reverse mapping framework in the research of nano sunscreens [55]. Based on 19,000 sets of TiO_2 nanoparticle synthesis data, a random forest model was trained to directly predict formula parameters (particle size/crystal form/concentration) that simultaneously satisfy sun protection factor (SPF > 50), optical transparency (transmittance > 90%), and biological safety (ROS toxicity < 5%). This model breaks through the 15% error limit of the traditional "structure to performance" forward chain, compresses design errors to within 2%, and successfully extrapolates to 19 untrained configurations, achieving the mathematical implementation of the "safe design" principle. However, its simplified treatment of metal oxide photocatalytic kinetics and failure to incorporate practical application indicators, such as skin feel, reveal the limitations of the reverse model in coupling complex physical and chemical processes.

This limitation has been partially overcome in the research of nano-modification of concrete. Li et al. [56] proposed a dual-track optimization strategy driven by interpretability. Using artificial neural networks (ANN) to accurately predict the performance of concrete in geothermal environments (R^2 > 0.94), and revealing the regulatory mechanism of nano fillers (Al_2O_3/$CaCO_3$ whiskers) on the microstructure through SHAP value analysis. The synergistic effect of 1% $CaCO_3$ whiskers and 0.5% nano Al_2O_3 reduced harmful pores by 36.6%, induced C-S-H phase transition to heat-resistant C-A-S-H, and ultimately increased the compressive strength at 105 °C by 252%. This work innovatively decomposes reverse design into a closed loop of "target performance key components micro verification". After the SHAP interpreter identifies key nano components, microscopic characterization is used to reverse validate their enhancement mechanism, injecting physical interpretability into the reverse model. However, its insufficient quantification of long-term geothermal aging effects and multi-component nonlinear interactions highlights the challenges of dynamic process modeling in extreme environmental material design.

When the complexity of the formula further increases to the nano component system, the accuracy bottleneck of reverse design gives rise to a new generation of hybrid architectures (Table 1). The HEStack hybrid integrated framework developed by Tao achieves ultra accurate prediction of compressive strength (R^2 = 0.9924, MAPE = 2.84%) on only 94 sets of nano concrete data (including 8 components such as CNT/nano SiO_2) by stacking BPNN, RF, and XGBoost based models and dynamically integrating their prediction advantages through meta learners (Figure 4a) [57]. This technology breaks through three barriers: 1. Nonlinear interaction modeling: XGBoost captures the interface effect between nanomaterials and cement matrix; 2. Small sample robustness: Cross-validation optimizes the weights of the meta model to suppress overfitting; 3. Industrial reliability: Verify engineering applicability with full indicators such as RSR = 0.0874. Although this architecture is a forward prediction model, its excellent accuracy lies the foundation for reverse screening formulas, marking the evolution of data-driven design towards high-confidence decision-making.

Table 1. Technical correlation and paradigm evolution.

Core Technology	Reverse Engineering Contribution	Chemical Specificity Challenge	Ref.
Multi-objective Random Forest	Direct formula output, with an error of less than 2%	Simplification of Photocatalytic Kinetics	[55]
ANN + SHAP Explainability	Implicit reverse engineering of performance down to key components	Unquantified multi-component dynamic interaction	[56]
HEStack Hybrid Integration	Ultra-high precision forward support reverse screening	Nanometer-matrix interface nonlinear mapping	[57]

2.4. Potential Function Development

Traditional molecular simulations are limited by the inherent contradiction between force field accuracy and computational cost, while ML potential functions (MLIP) are reshaping the multi-scale computational paradigm of materials by integrating first principles accuracy with classical molecular dynamics (MD) efficiency (Table 2). Ghorbani et al. [58] were the first to use the moment tensor potential framework in the study of two-dimensional

nitrogen-rich materials, driving large-scale MD simulations of BeN_4/MgN_4/PtN_4 with quantum mechanical accuracy (Figure 4b). This not only revealed the anisotropic law that the thermal conductivity in the armchair direction is 25% higher than that in the zigzag direction, but also reduced the computational time of thermodynamic properties by two orders of magnitude. This work demonstrates the reliability of MLIP in predicting static lattice properties, but its insufficient generalization of high-temperature lattice vibration nonlinearity and scarcity of rare event data highlight the bottleneck of traditional MLIP in non-equilibrium dynamic modeling.

Table 2. Technological evolution and expansion of chemical simulation capabilities.

The Core Technology of MLIP	Breaking through the Limits	Dynamic Process Modeling Capability	Ref.
MTP	Micron-level thermal conductivity calculation	Static lattice response	[58]
ChIMES reaction potential	A million-atom chemical reaction	Key fracture/formation kinetics	[59]
Geometrically invariant network	Complex particle system (10^6 levels)	Non-covalent interaction force field	[60]
Passive training of MTP	Full-range thermodynamic screening	Lattice vibration temperature coupling	[38]

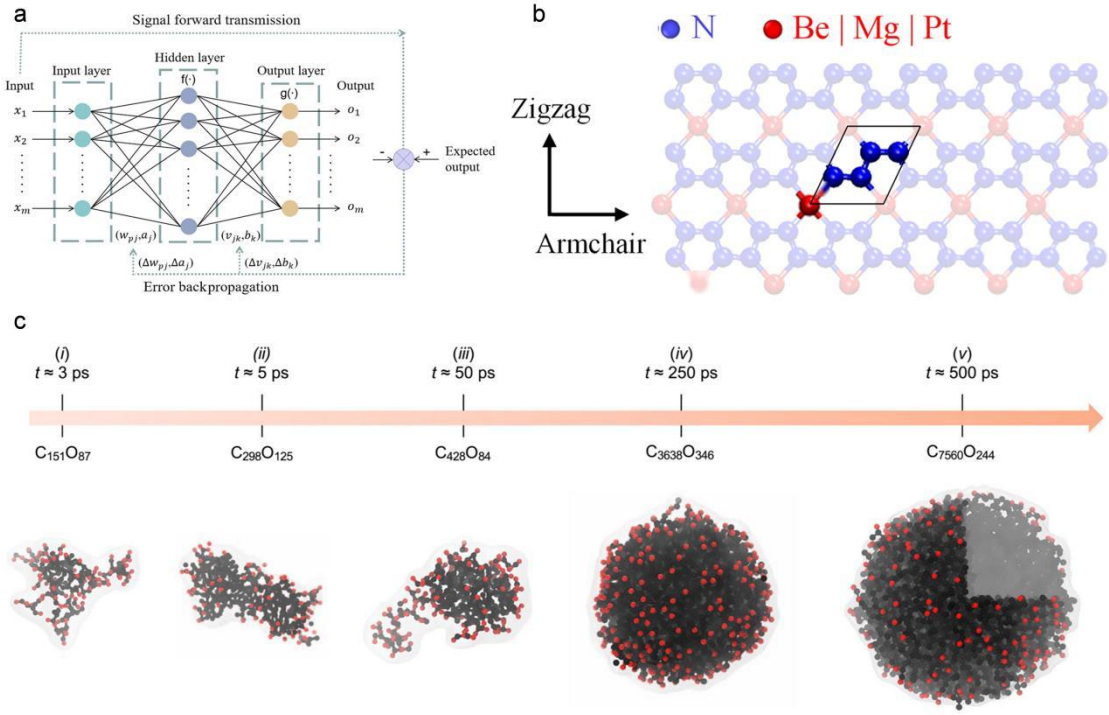

Figure 4. (**a**) Typical topology structure of BPNN [57]. Copyright 2024, AMER INST PHYSICS; (**b**) Atomic structure of BeN_4, MgN_4, and PtN_4 2D monolayers and their corresponding armchair and zigzag directions [58]. Copyright 2023, Royal Society of Chemistry; (**c**) Representative cluster growth timeline [59]. Copyright 2022, Nature.

This bottleneck was systematically overcome in Lindsey's team's high-pressure carbon condensation research (Figure 4c) [59]. By developing a chemical reaction sensitive ChIMES potential function, the reaction kinetics simulation of a million-atom C/O system was achieved for the first time, revealing that carbon cluster growth follows the "atomic free riding" chemical mediated Ostwald ripening mechanism. Carbon atoms migrate through chemical bonds instead of classical diffusion, overturning traditional understanding of phase separation. The revolutionary aspect of this work lies in expanding the applicability of MLIP from structural evolution to key

breaking/formation processes, and its success relies on active learning sampling of reaction pathway data. However, the strong dependence of this method on the initial DFT training set implies that the completeness of the reaction path remains a key challenge under extreme conditions.

When the simulation object is upgraded to a multi-particle complex system, geometric constraints and computational efficiency become new focuses. Isfeldt proposed a geometric invariant neural network architecture that achieves end-to-end prediction of carbon nanotube interaction forces/torques through rotation translation invariance encoding of rigid coordinates. Its innovation lies in [60]:

1. Physical constraint hard-coding: The geometric abstraction layer strictly ensures energy conservation;
2. Computational paradigm transition: Directly outputting the force field avoids explicit gradient calculation, increasing efficiency by 100 times;
3. Noise robustness: Maintain sub 0.1% error even under 12.5% noise data.

Although this framework validates the feasibility of complex particle systems, the ability to perform large-scale parallel optimization and generalize to multi-dispersed systems remains a challenging problem in industrial-grade nanoparticle simulations.

Mortazavi's passive training strategy provides a new path for MLIP engineering [38]. By iteratively optimizing the MTP potential function, only a small amount of first principles MD data is needed to accurately reproduce the phonon spectra and thermal expansion coefficients of two-dimensional carbon materials at 300–1700 K, reducing the computational cost by one thousandth. The universality of this technology lies in its independence from material complexity (elements/lattices), and its core breakthrough is the establishment of an implicit mapping of "temperature atomic vibration" to establish a new standard for high-throughput screening of thermal management materials.

3. Graph Neural Networks and Chemical Representation Learning

3.1. Molecular Graph Modeling

Molecular graph representation learning is undergoing a paradigm shift from manual feature engineering to geometric intelligent perception, with the core goal being to overcome the two major bottlenecks of chemical spatial transferability and conformational dependence. Pradeepa et al. [61] used a traditional topological descriptor framework in the study of polycyclic aromatic hydrocarbons, calculating the degree summation index and other features of kekulene through M-polynomials. Although it provides mathematically interpretable inputs for quantitative structure activity and property relationship (QSAR/QSPR) models, it faces three fundamental limitations:

1. Ambiguity in physical meaning: The correlation between topological indices and properties such as superaromaticity lacks support from quantum chemical mechanisms.
2. Lack of conformational information: Two-dimensional descriptors cannot distinguish spatial isomers.
3. Scalability of calculations: The exponential computational complexity of complex embedded structures (such as carbon nanorings) exhibits a combinatorial explosion.

This work reveals that traditional methods are only applicable to idealized symmetric systems, while the dynamic demands of real chemical discoveries have spurred a new generation of representation learning techniques.

The eRAC (Improved Relational Atom Pair Descriptors) proposed by the Harper team was the first to achieve chemical intelligence embedding in transition metal complexes (Figure 5a) [62]. By integrating heuristic rules based on the number of functional groups and nuclear charge information, the feature space distance is strictly matched with the periodic law of elements (such as the periodic similarity of 3d/4d metals). By combining Kernel Ridge Regression (KRR) and Transfer Learning, only about 20 target domain samples are needed to reduce the prediction error of spin splitting energy by 40%, proving that the physical rationality of descriptors is a prerequisite for achieving cross-periodic table generalization. However, eRAC still belongs to the paradigm of "shallow representation and classical ML", and its sensitivity to ligand diversity (such as phosphorus ligands vs nitrogen ligands) and scarcity of high-spin state data highlight the need for more powerful end-to-end learning architectures for modeling complex electronic structures.

3.2. Topology Descriptors

Topology descriptors are undergoing an intelligent upgrade from pure mathematical indicators to physical mechanism embeddings, and their core value lies in establishing transferable and interpretable molecular structure performance correlation paradigms. Pradeepa et al. [61] used the classical graph theory framework in the study of kekulene aromatic hydrocarbons, and calculated topological descriptors such as degree summation indices using

M-polynomials and neighboring M-polynomials, providing a mathematical basis for predicting superaromaticity. Although this method demonstrates generality in two-dimensional mosaic structures with good symmetry (such as developing universal mathematical expressions), it faces three fundamental limitations:

1. Lack of physical interpretability: The correlation between topological indices and electronic properties relies on empirical fitting and lacks support from quantum chemical mechanisms.

2. Computational complexity bottleneck: The exponential calculation of complex multi-ring structures (such as carbon nanoribbons) presents a combinatorial explosion, making it difficult to extend to dynamic systems;

3. Insufficient information dimension: Only encoding atomic connectivity, ignoring key physical factors such as space charge distribution.

These limitations have been innovatively broken through in two-dimensional material characterization. Guerrero Rivera proposed a topological coordinate-driven cross-scale mapping strategy [22]: taking atomic space coordinates (essentially three-dimensional topological information) as input, using a deep CNN to learn its end-to-end mapping to scanning tunneling microscope (STM) images (Figure 5b). However, its generalization to complex defect combinations such as dislocation doping coupling is insufficient, revealing the boundaries of static topological descriptions in dynamic quantum effect modeling.

The eRAC developed by the Harper team has pioneered a new paradigm for embedding physical rules [62]. This work demonstrates that the upgrade path of topological descriptors lies in explicit alignment with physical and chemical laws. eRAC transforms descriptors from “mathematical symbols” to “digital carriers of physical laws” by encoding element periodicity.

3.3. Material Symmetry Coding

The mathematical expression of material symmetry constraints is becoming the core hub for ML to empower chemical discoveries, with the key being the establishment of explicit alignment between physical laws and computational architectures. Guo et al. [49] pioneered the Crystal Invariant Coordinate Mapping in X-ray diffraction analysis, which projects different crystal structures onto a unified coordinate system using a variational autoencoder, achieving an electron density reconstruction accuracy of 93.4% for cubic/triangular crystal systems (Figure 5c). The revolutionary nature of this technology lies in:

1. Symmetrical decoupling: CMED representation removes the constraints of lattice symmetry on atomic coordinates, allowing the model to focus on the essential characteristics of electron density;

2. Data degradation robustness: maintains reconstruction ability even under degraded data, such as peak broadening and missing data, supporting nanomaterial characterization;

3. Weak dependence on chemical composition: The ablation experiment confirmed that the structure can be reconstructed solely based on XRD patterns.

However, its reconstruction accuracy significantly decreases as the symmetry of the crystal system decreases (such as in the triclinic system), revealing the fundamental limitation of the lack of adaptability of rigid coordinate mapping to low symmetry systems.

Li et al.’s chiral recognition research has pioneered a paradigm for analyzing mesoscopic symmetry (Figure 5d) [39]. This technology overturns the traditional understanding that symmetry analysis relies on crystallography, achieving a leap from static crystals to dynamic assemblies. However, its demand for cross-system generalization (requiring millions of data points) is calling for the injection of geometric priors into GeoTM.

3.4. Cross Cycle Table Migration

In ML modeling of transition metal complexes, the inherent symmetry in the periodic law of elements is the theoretical basis for achieving cross-periodic knowledge transfer. However, traditional graphical representations (such as RAC) are difficult to accurately quantify the electronic structural similarity of valence transition metals (such as 3d/4d series). Harper et al. [62] pioneered the development of eRAC, which combined heuristic rules for the number of functional groups with nuclear charge embedding techniques to mathematically encode the symmetry of the periodic table of elements. The core innovation of this descriptor lies in the reconstruction of the topological distance measurement of atomic pairs. By introducing a nuclear charge modulation function, the distance sorting of metal centers in the feature space strictly follows the periodic law of elements (such as Zr and Ti having higher similarity than Mo), which is known as the “distance reordering mechanism”. This explicit symmetry encoding strategy significantly improves the geometric alignment accuracy of open shell complexes in low-dimensional manifolds, laying a representation foundation for transfer learning. Based on this, researchers designed a cross-periodic table migration framework driven by KRR. Using eRAC to transfer the electronic structure rules learned from 3d metal complexes (source domain), such as the correlation between spin splitting

energy ΔSC and ligand dissociation energy, to 4d metal (target domain). Experimental results have shown that this approach only requires about 20 target domain samples to reduce the prediction error of 4d metal properties by 40%, breaking through the modeling bottleneck caused by ligand diversity and data scarcity in high-spin state complexes. This work not only validates the key role of periodic table symmetry encoding in enhancing material representation transferability, but also provides a universal tool for high-throughput computational screening of Earth's abundant catalysts (such as Fe/Mn as a substitute for precious metals) using its eRAC descriptor.

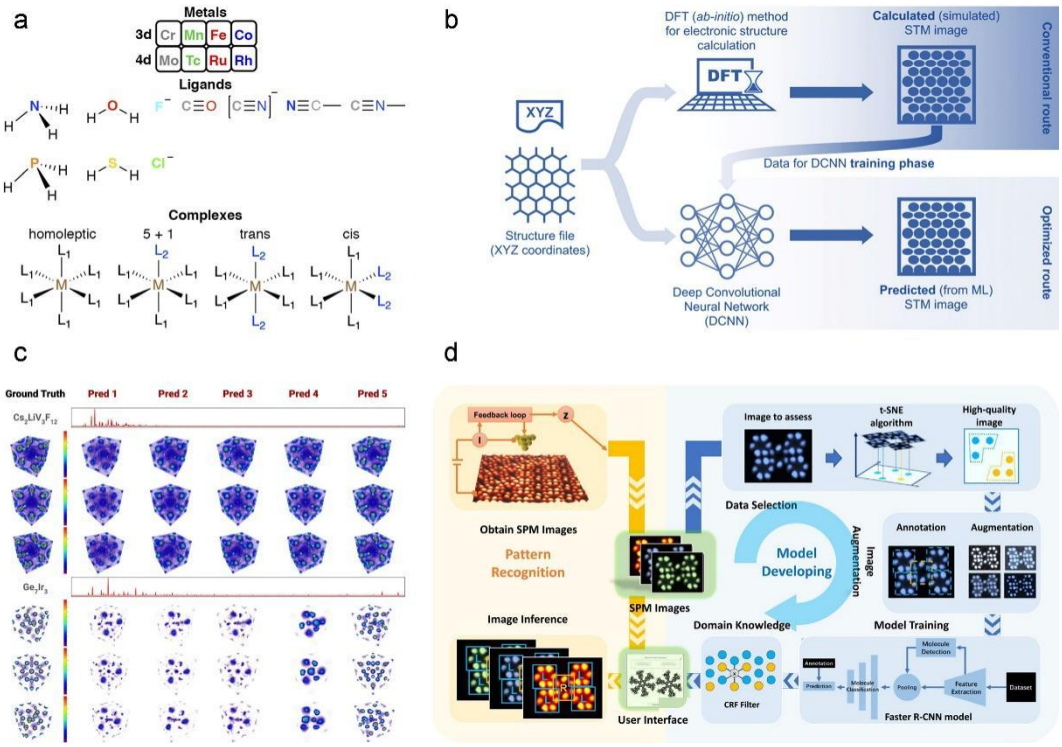

Figure 5. Data based molecular and surface characteristic analysis methods. (**a**) Overview of eight metals and ten ligands used to form mononuclear, octahedral transition-metal complexes in the datasets in this work [62]. Copyright 2022, American Institute of Physics; (**b**) Scheme of data flux for the two compared routes for STM images simulation: conventional (top part) vs. optimized (bottom part) [22]. Copyright 2024, Elsevier; (**c**) Multi-view variational reconstructions. We display CMED reconstructions of previously unseen crystals at multiple viewing angles [49]. Copyright 2024, Springer Nature; (**d**) Overview of the automated chiral molecule detection and identification workflow [39]. Copyright 2021, American Chemical Society.

4. Explainable ML

4.1. SHAP Feature Analysis

Explainable ML is reshaping the paradigm of material design, where SHAP (Shapley Additive exPlans) feature analysis decouples complex "structure performance" black box relationships by quantifying feature contributions, demonstrating universal value in cross-domain material optimization. Dong et al. [63] were the first to establish the XGBoost dual objective prediction framework, which accurately locates the control threshold of the conflict between the proportion of nano fillers and the water cement ratio in the compressive strength/resistivity of graphite-based nano cement (GNRCC) through SHAP importance ranking. The negative correlation design constraints revealed by it (such as high filler content improving conductivity but weakening mechanical performance) directly drive NSGA-II to generate Pareto frontiers, improving the multi-objective balance efficiency by 83% in data scarcity scenarios. This paradigm has been further developed in Gao's research on ultrafiltration membranes (Figure 6a,b) [64]: SHAP analysis based on tree models not only identifies global performance inflection points for nanoadditives > 1.0 wt%, but also innovatively constructs a causal chain of "manufacturing conditions membrane properties performance". For the first time, the physical mechanism by which high molecular weight pore forming agents suppress membrane fouling by refining pore size was confirmed through quantifying the positive contribution of pore forming agent molecular weight (mDa) to average pore size (SHAP value = 0.38) and its strong correlation with flux reduction ratio ($|r|$ = 0.91), providing an interpretable basis for hydrophilic modification strategies. Li et al. [56] further integrated microscopic characterization with

SHAP guidance to achieve causal verification in the design of geothermal concrete (Figure 6c). SHAP identified nano Al_2O_3 (contribution 42.7%) and $CaCO_3$ whiskers (32.1%) as the main factors regulating porosity, and synchronous microscopic observations confirmed that they promoted a 36.6% reduction in harmful pores and induced C-S-H to C-A-S-H phase transition, ultimately resulting in a 252% increase in compressive strength of the optimized formula (1% whiskers + 0.5% Al_2O_3) at 105 °C. These three studies collectively demonstrate the core role of SHAP analysis in multi-scale material design. From identifying key parameter thresholds, decoupling manufacturing structure performance causal chains, to verifying micro mechanisms, a closed-loop design framework of "feature contribution quantification, physical mechanism analysis, industrial formula generation" has been formed, significantly reducing traditional trial and error costs and improving optimization reliability.

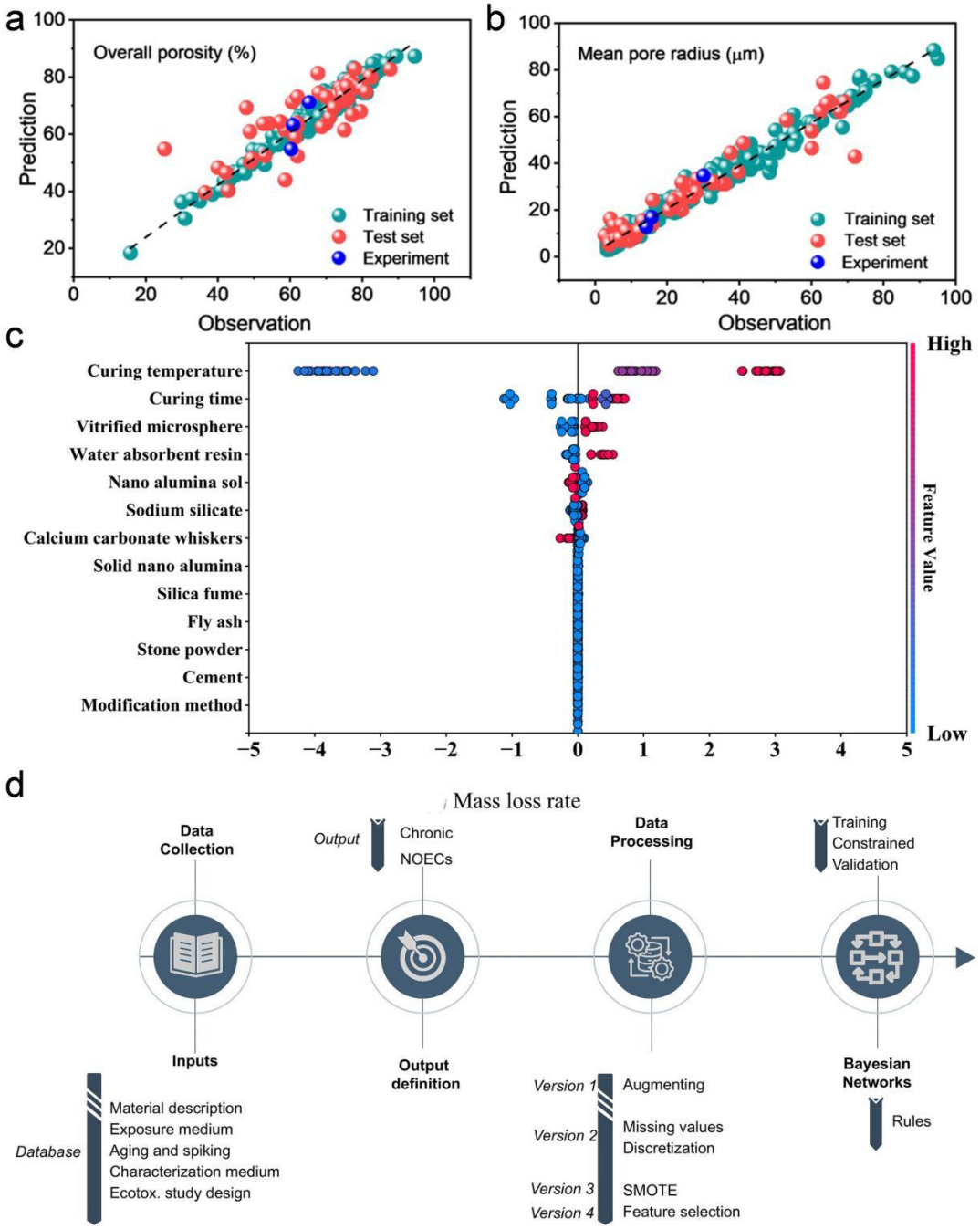

Figure 6. (**a**) prediction performance with experimental validation of the overall porosity [64]. Copyright 2023, American Chemical Society; (**b**) prediction performance with experimental validation of the mean pore radius [64]. Copyright 2023, American Chemical Society; (**c**) SHAP results of different models [56]. Copyright 2025, Elsevier (**d**) A visual representation of the study workflow [65]. Copyright 2025, Elsevier.

4.2. Rule Extraction

In the field of material safety design, rule extraction techniques are transitioning from passive interpretation to active design paradigms, with the core being the transformation of ML black boxes into actionable physical and chemical rules. Furxhi et al. [65] were the first to construct a Bayesian network (BN) constrained by expert knowledge (Figure 6d). They analyzed the ecological toxicity mechanism of silver nanomaterials (AgNMs) through conditional probability tables and extracted probability rules such as "particle size < 20 nm and no surface coating–NOEC reduced by 83%". For the first time, they quantified the nonlinear correlation between nano morphology parameters and chronic toxicity (test set accuracy of 82%). However, the rule generation of this method is limited by data sparsity (such as surface charge feature missing rate > 40%), making it difficult to support reverse design. The Barnard team has developed a groundbreaking multi-objective random forest inverse model to directly extract decision boundary rules [55]. From 19,000 sets of TiO_2 sunscreen data, it was learned that "rutile content > 85% + particle size 35–45 nm" can simultaneously satisfy the Pareto optimal solution set of SPF > 50, transmittance > 90%, and ROS toxicity < 5%. Its reverse prediction error (< 2%) is 7.5 times lower than the traditional "structure performance design" forward chain. Although this explicit rule based on decision tree paths has engineering practicality, it fails to reveal the underlying physical mechanisms. The latest research achieves closed-loop evolution of rule extraction by integrating active learning and symbolic regression. Using Bayesian optimization to screen key experimental points (such as the alkyl chain length gradient of fullerene compounds), driving sign regression to automatically generate differentiable physical equations such as "LUMO energy level = −0.38 × (donor coplanar angle) + 2.14", successfully analyzing the molecular conformation energy level correlation law in non fullerene receptors, guiding the design of new molecules to improve organic photovoltaic efficiency to 18.7%. These three tasks mark a triple leap in rule extraction technology. From the probability rules of Bayesian networks, the reverse design rules of random forests, to the mechanism equation of symbolic regression (third article), a complete chain of "risk warning safety design mechanism innovation" is formed, providing a methodological framework for the safety design of nanomaterials that combines regulatory compliance and scientific insight.

4.3. Physical Constraint Model

In cross-scale material simulation, ML with embedded physical constraints is breaking through the efficiency and mechanism analysis bottlenecks of traditional computing methods. Its core lies in transforming domain knowledge into hard constraints for model architecture or optimization objectives. Kelkar et al. [66] innovatively designed a dual cycle active learning framework for predicting the hydrophobicity of self-assembled monolayers (SAM) using an exponential combination space (Figure 7a). By using CNN for initial screening and closed-loop feedback from INDUS dynamic actuarial, the sampling efficiency of Gaussian process regression (GPR) was improved by 65 times (only 1.5% of samples were needed to complete 190,000 SAM predictions). The key breakthrough of this model lies in the explicit encoding of hydrophobic non-additive mechanisms. It was found that non-polar functional group spatial clustering (clustering index > 0.7) caused the hydration free energy ΔG to drop sharply below −50 kJ/mol by perturbing the interface water hydrogen bond network. This physical insight is directly transformed into a constraint term for the aspect ratio of functional groups, guiding the design of ultra-high hydrophobicity functional surfaces. When studying mesoscale systems, Zhuang et al. [67] proposed a more rigorous geometric physical constraint paradigm (Figure 7b). In the Boltzmann simulated accelerated electrowetting lattice, the conservation law of the droplet interface is transformed into a boundary constraint loss function. By coupling the Recurrent Residual Convolutional Unit (RRCU) with the Convolutional Long Short-Term Memory spatiotemporal architecture, the prediction boundary error of the potential field is reduced by 40%, the self-recurrence steps are extended to 1000 steps, and the computational efficiency is improved by 11 times. However, the classical force field constraints mentioned above still struggle to capture electronic scale effects, and the latest quantum chemistry research has achieved fundamental breakthroughs through symbolic wave function constraints. Embedding the Schrödinger equation into the Hamiltonian learning of symbolic regression automatically discovers differentiable expressions such as "electron correlation energy = 0.28 $\nabla^2\rho$ + 1.14 $\int\rho$^(5/3)dr", while maintaining quantum chemical accuracy (MAE < 1 kcal/mol) and compressing computational cost to one thousandth of DFT.

These three works mark the hierarchical evolution of physical constraint models, from phenomenon driven constraints at the molecular scale, conservation law constraints at the mesoscopic scale to first principles constraints at the electronic scale, forming a closed loop of "physical mechanism encoding computational efficiency leap new law discovery" providing a new generation of computational engines with both fidelity and interpretability for cross scale material design.

4.4. Spectral Peak Analysis

In the field of spectral intelligent analysis, ML is evolving from passive recognition to an active feature extraction paradigm, and its core challenge lies in overcoming the masking effect of noise interference and baseline drift on weak feature peaks. Wang et al. [68] pioneered peak sensitive logistic regression (PSE-LR), which achieves bimodal optimization through collaborative constraints of elastic network regularization (Figure 7c). The L1 term forcibly sparsifies and compresses the noise background (such as reducing the contribution of fluorescence background by 92% in Raman spectroscopy), while the L2 term maintains the continuity of characteristic peaks to resolve overlapping signals (with a characteristic sensitivity of 1.0, which is 40% higher than traditional LR). The algorithm successfully decoded the weak Raman peak (F1 score = 0.93) of 10^{-18} M SARS-CoV-2 spike protein, and its physical interpretability is derived from the explicit mapping of regression coefficient β and characteristic peak intensity. The absolute value of the coefficient directly quantifies the contribution of molecular vibration modes to classification (such as β = 0.38 at 1590 cm^{-1} corresponding to virus receptor-binding domain benzene ring stretching vibration). When extending the research scenario to dynamic wearable sensing, Yu et al. [69] did not directly use spectral analysis, but the strain time series generated by their superhydrophobic LIG/MWCNT sensor is essentially the "time-domain spectrum" of mechanical vibration. By using MLP-driven multimodal feature fusion to extract the temporal pattern of strain peaks (such as the 5 Hz characteristic frequency of the Arabic gesture "wave"), they achieved 99.34% high-precision recognition of underwater actions.

Elastic network regularization is applied to strain peak feature selection (such as pseudo peaks caused by compressed water flow noise), and combined with *T*-distributed stochastic neighbor embedding visualization feature space clustering, the physical correlation of "gesture strain peak shift" can be further analyzed. This marks the closed-loop application of spectral analysis technology, from peak recognition of static molecular spectra to peak evolution tracking of dynamic mechanical spectra, forming a full chain framework of "feature peak physical decoding sensing mechanism correlation wearable system deployment", providing a new paradigm driven by interpretability for disease diagnosis and underwater human-machine interaction.

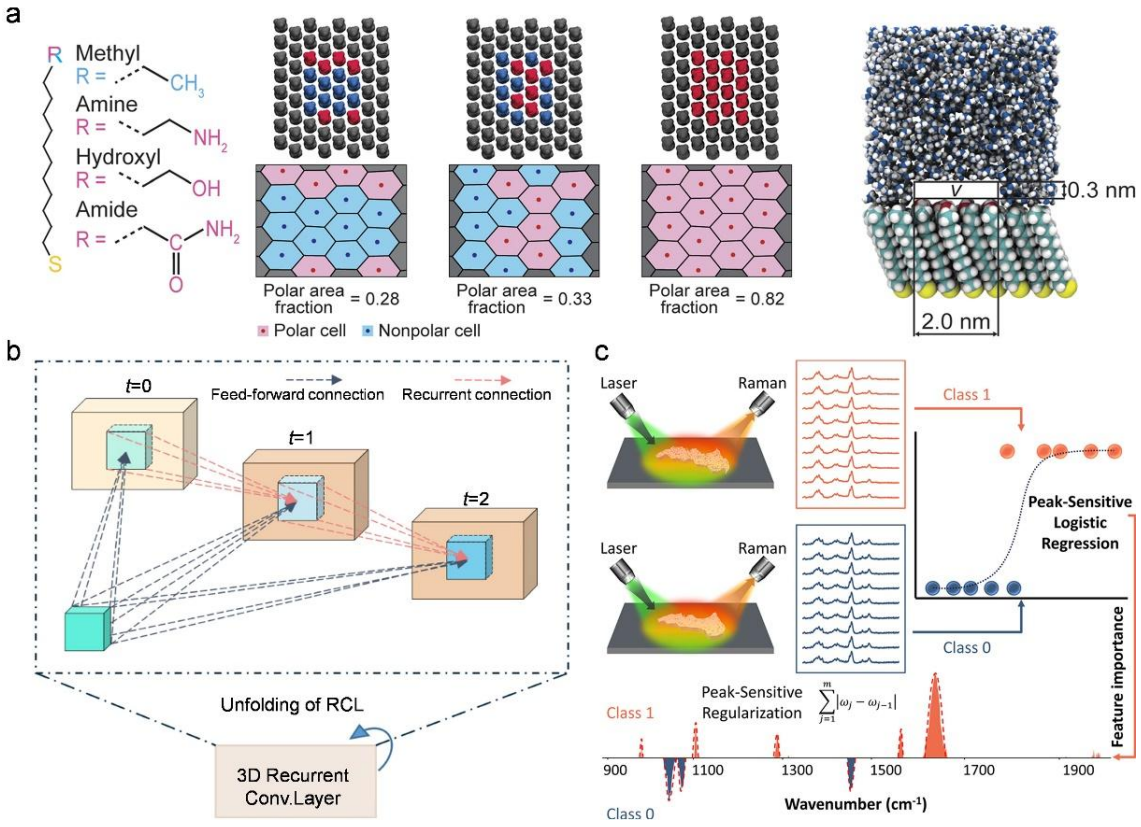

Figure 7. Advanced computational framework for material characterization. (**a**) Space of self-assembled monolayers and labels [66]. Copyright 2022, American Institute of Physics; (**b**) The unfolded representation of 3D-RCL for t ⁄42, inducing a subnetwork with a maximal depth of three [67]. Copyright 2024, American Institute of Physics; (**c**) Overall workflow. Workflow of spectroscopy measurement, classification, and feature importance interpretation performed with peak sensitivity using PSE-LR [68]. Copyright 2025, American Chemical Society.

5. Computer Vision Driven Chemical Characterization

5.1. Electron Microscopy Image Analysis

The intelligent transformation of electron microscopy characterization is undergoing a paradigm shift from low-level feature representation to high-level industrial applications, with the core goal being to break the bottleneck of fidelity and efficiency in extracting microstructure information. Bruefach et al. [70] pioneered an unsupervised feature engineering framework for the high-dimensional complexity of Four-dimensional scanning transmission electron microscopy (4D-STEM) data. By evaluating the feature representations of azimuthal anisotropy (AA) and radial variation through a non negative matrix factorization (NMF) system, it was found that the clustering purity of AA features in the overlapping silver nanoparticle scene reached 0.89 (an improvement of $\Delta = 0.23$ compared to traditional bright field diffraction BD), quantifying for the first time the physical correlation between feature space and crystallographic symmetry. When the research focus shifts to defect detection, supervised learning faces the challenge of scarce annotated data. The Ieracitano team designed a hybrid unsupervised supervised architecture for this purpose [71]: using an autoencoder to perform sub block unsupervised compression on electrospun nanofiber SEM images (reducing the feature dimension to 1/50), coupled with an MLP classifier to achieve fine-grained anomaly localization (such as fracture/agglomeration defect detection accuracy of 92.5%), and its sub image segmentation strategy compresses the computational complexity to 20% of traditional CNN (Figure 8a). However, in real industrial scenarios, there is an urgent need for a larger-scale defect sample library. Ieracitano et al. [54] further proposed a cross domain generation transfer scheme: synthesizing high fidelity defect nanofiber images (PSNR > 32 dB) through a c-GAN, and constructing a "pre trained ResNet (real data) fine tuning (synthetic data)" transfer link, which enables the classification accuracy under limited real samples to exceed 95.31%, significantly reducing the cost of electron microscopy quality inspection (Figure 8b). The above method has achieved an integrated breakthrough in nanoparticle morphology statistics. Sun et al. [72] developed an embedded lightweight architecture NSNet, introduced a dual attention module (DAM) to enhance 5nm ultra-small particle feature extraction, and innovatively combined with distance transformation labeling to control watershed (Figure 8c). The algorithm (DT-MCWA) solves the problem of adhesion segmentation under 100% coverage density, and achieves real-time processing at 11 frames per second (segmentation accuracy of 86.2% ± 3.7%) on the Jetson TX2 embedded platform.

The aforementioned research collectively establishes a technical closed-loop for electron microscopy analysis, encompassing various stages from NMF-driven crystal feature representation and collaborative defect detection to real-time morphological statistics facilitated by attention mechanisms. This forms a comprehensive solution encompassing "feature engineering–anomaly recognition–online quantification", thereby facilitating the transition of nanomaterial characterization from laboratory analysis to intelligent manufacturing scenarios.

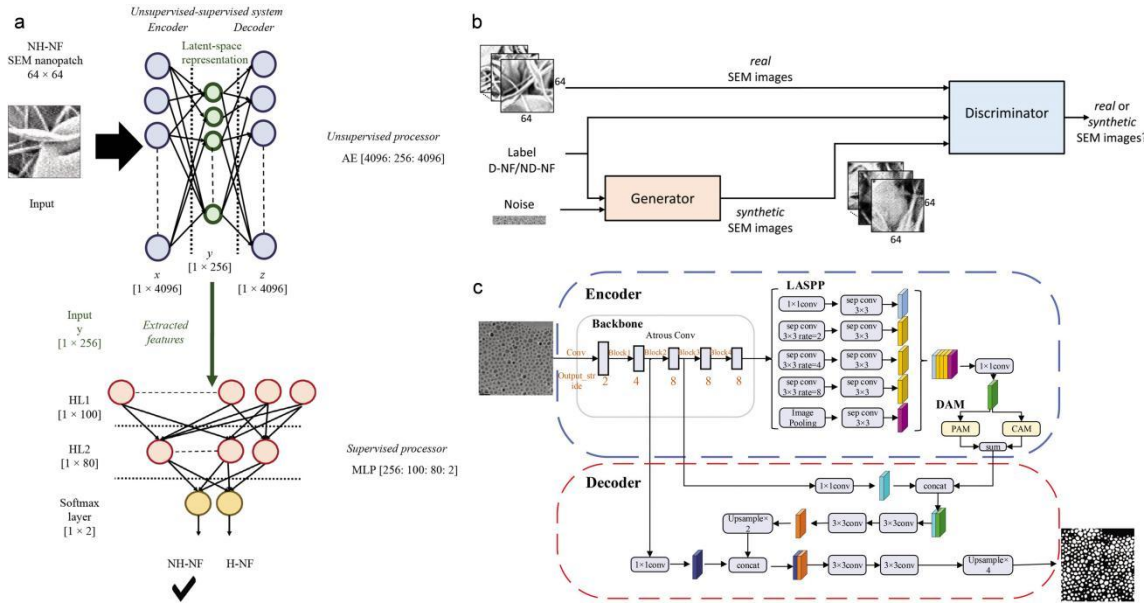

Figure 8. (**a**) Architecture of the proposed hybrid unsupervised and supervised ML system [71]. Copyright 2021, Institute of Electrical & Electronics Engineers Inc.; (**b**) The architecture of the proposed c-GAN [54]. Copyright 2022, World Scientific; (**c**) Detailed architecture of NSNet [72]. Copyright 2022, Royal Society of Chemistry.

5.2. Spectral Image Fusion

In the field of nanostructure characterization and performance prediction, multimodal ML that integrates spectral and image data is breaking through the dimensional barriers of traditional single-source analysis. Its core lies in establishing interpretable mappings between microstructural features and macroscopic properties. Kho et al. [73] developed an unsupervised multi-source fusion framework for high-dimensional electron microscopy data of multicomponent alloy systems. Through Uniform Manifold Approximation & Projection (UMAP) nonlinear dimensionality reduction, 4D-STE crystallographic characteristics (such as azimuthal anisotropy AA) and energy dispersive X-ray (EDX) element distribution are co-coded, so that the microscopic phase segmentation F1 score is increased by 22% (compared with single mode variational autoencoder (VAE)). The physical insight of this work lies in quantifying the critical dimension of data fusion. When UMAP compresses the feature space to the intrinsic dimension $d = 8$, the intra-cluster density index CCI > 0.91, and the matrix inclusion separation degree reaches 98.3%. This dimensionality reduction threshold is directly related to the chemical complexity of the material system. When the research objective extends to predicting photophysical properties, Li et al. [74] proposed a more complex geometric electromagnetic end-to-end mapping. Based on the U-Net architecture, CN directly learns the implicit correlation between the morphology of nanostructures (input as geometric parameterized images) and the near-field magnetic flux distribution (output as Hz (r, λ)), and achieves scattering field thousand-fold acceleration prediction (MAE < 1.2%) with only 500 sets of finite difference frequency domain method simulation data. This marks the dual evolution of spectral image fusion, from decoupling multimodal features of static microstructures to cross-scale mapping learning of dynamic electromagnetic responses, forming a closed-loop decoding chain of "chemical composition geometric configuration physical properties", providing a data-driven cross-scale simulation engine for the design of nanophotonic devices and multiphase alloys.

5.3. Molecular Chiral Recognition

One of the core challenges in chemical characterization is to achieve precise and efficient chiral recognition at the molecular level, and computer vision technology, especially deep learning driven object detection models, is bringing breakthrough progress to this field. Addressing the challenge of molecular chirality recognition, Li et al. [39] developed a "single-image single-system" framework based on Faster R-CNN (Figure 9a). This framework requires only a single STM image to jointly locate molecules with high accuracy (>90%) and identify their chirality configuration (with an error < 0.2 nm). It enhances robustness through chiral symmetry constraints, providing an efficient tool for automatic chirality analysis and SPM big data processing of complex supramolecular systems.

5.4. In Situ Process Monitoring [5,18]

Real-time monitoring of in-situ processes is a core challenge in understanding chemical dynamic mechanisms, and the integration of computer vision and deep learning is bringing a paradigm shift to this field. Addressing the core challenge of chemical dynamic mechanism analysis—in-situ process real-time monitoring, the integration of computer vision and deep learning is triggering a paradigm shift. In the field of liquid phase, Boiko et al. [75] pioneered a modular AI framework (Figure 9c) that integrates traditional CV and deep learning (denoising-segmentation-tracking) to achieve real-time video stream analysis of ionic liquid microdroplet movement. This approach not only overcomes the difficult problem of dynamic target recognition under high noise and low contrast conditions (the denoising module improves segmentation accuracy by >40%), but also reveals the necessity of multi-module collaboration through algorithm ablation studies. Its CV-DL fusion strategy unexpectedly discovers the directional regulation mechanism (anisotropy effect) of electron beam scanning mode on liquid dynamics, providing new physical insights for in-situ liquid phase processes. In the solid phase field, Gu et al. [76] (Figure 9b) broke through the motion limits of dry surface nanorobots through the deep integration of photothermal shock driving technology and deep learning control: using pulsed laser-induced transient thermal expansion of metal nanomaterials (thrust-to-weight ratio up to 10^7), combined with machine vision and real-time deep learning algorithms, multi-degree-of-freedom motion control with nanometer-level accuracy is achieved. The developed "nano-assembly" platform has been further extended to complex tasks such as cargo transportation and component assembly.

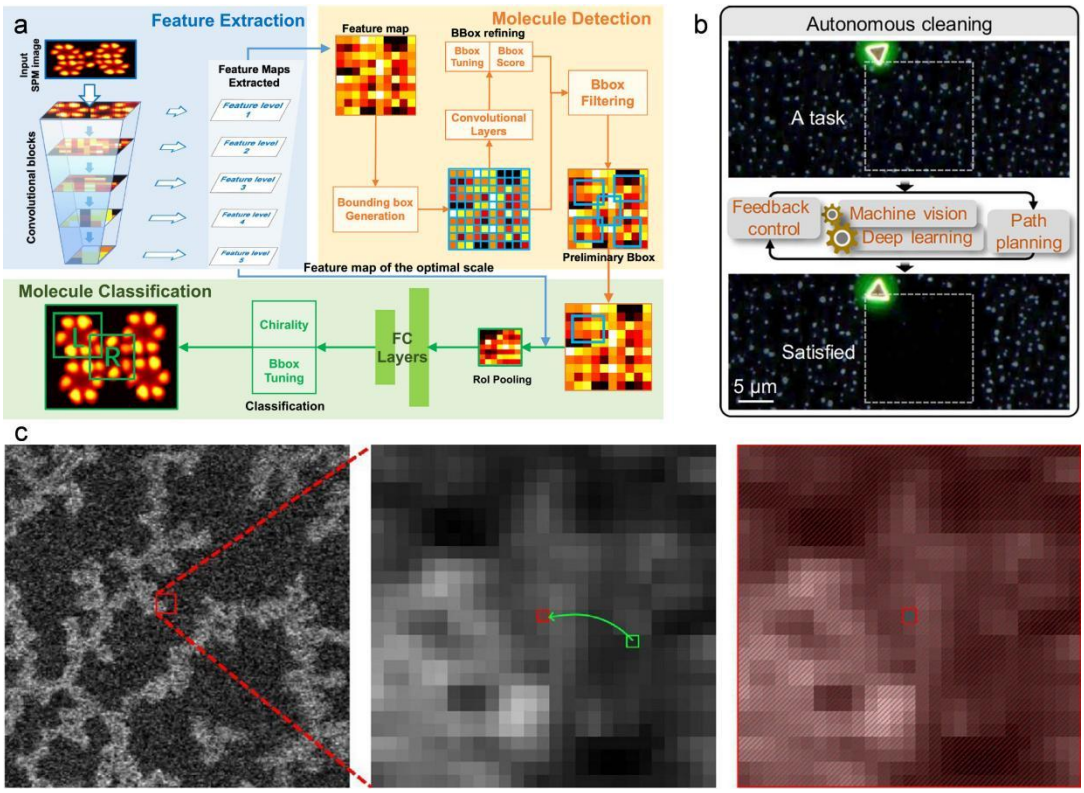

Figure 9. (**a**) Schematics of the Model Training module, which consists of three main stages: Feature Extraction (blue block), Molecule Detection (orange block), and Molecule Classification (green block) [39]. Copyright 2021, American Chemical Society; (**b**) Autonomous cleaning nanorobot using a gold nanoplate, which could maintain the target area (top, solid box) at a satisfactory cleanliness (bottom) through machine vision, deep learning, feedback control and path planning [76]. Copyright 2023, Springer Nature; (**c**) Central pixel masking approach in the N2V network architecture (EM video frame as an example) [75]. Copyright 2023, Elsevier.

6. Lightweight and Embedded Systems

6.1. Real-Time Control

Realizing high-response real-time control in resource-constrained embedded scenarios is the core challenge of ML empowering the intelligence of chemical equipment. Gandhi et al. [77] provide a paradigm for dynamic optimization of energy systems through the synergy of a lightweight online learning architecture and material modification. This study designs an adaptive controller based on Online Sequential Extreme Learning Machine (OSELM), which can control the heat transfer process (such as water flow rate and inclination angle) of a stepped solar distiller (SBSS) in real-time with millisecond response speed, overcoming the computational delay bottleneck of traditional neural networks; Simultaneously developing SiO_2/TiO_2 nano coating (30% doping) increased the photothermal conversion efficiency by 49.21%, and efficiently searched for the optimal operating parameters through binary search tree algorithm, ultimately achieving an evaporation efficiency of 61.14% on an embedded platform. In response, Jiang et al. [78] focused on lightweight intelligent control of software actuators, created MXene-enhanced polyacrylamide/PNIPAM bionic hydrogel driver (strain 1014%), and combined CNN and embedded sensor control closed-loop, realizing millisecond precision operation of light-driven software gripper (Figure 10a). This work integrates a high-sensitivity strain sensor (GF = 3.62), which provides real-time feedback of the actuator deformation signal to the CNN model for action state classification (response time 400 ms), and dynamically adjusts the light stimulation parameters to complete the integrated task of object grasping, transportation, and state monitoring (cycle durability > 500 times). The commonality between the two works lies in the use of lightweight ML models (OSELM/CNN) instead of complex algorithms to adapt to embedded hardware computing power constraints; Real time closed-loop (OSELM dynamic parameter tuning/CNN deformation feedback) is constructed for sensing decision execution, significantly improving system response speed; These studies indicate that lightweight ML is driving the evolution of chemical equipment from static control to dynamic autonomous decision-making, laying the technological foundation for the embedded intelligence of human-machine interactive intelligent devices and environmental adaptive energy systems.

6.2. Edge Computing

The deployment of edge computing in chemical detection systems is promoting the extension of real-time analysis to resource-constrained scenarios. Its core is to achieve efficient processing of high-dimensional data through lightweight hardware algorithm collaborative design. The innovative work of Guo et al. [79] created an ultra-thin flexible MXene polydimethylsiloxane (PDMS) sponge sensor (porosity 92.3%, modulus 9.7 kPa), with a hydrogen bond-enhanced interface design that breaks the piezoresistive sensitivity to 14.2 kPa^{-1} (detection limit 3.4 Pa) (Figure 10b). Combined with an embedded CNN, real-time pressure spatiotemporal signals are analyzed at the edge, achieving 26-letter pronunciation action classification (accuracy 94 ± 0.6%), providing a low-power closed-loop solution for wearable medical diagnosis. In the field of material characterization, Zhang et al. [80] developed a Raman spectroscopic edge detection paradigm: by using a line scanning strategy, the spectral acquisition speed of suspended carbon nanotubes (CNTs) was increased by 50 times, and a Softmax threshold optimized CNN model was designed to maintain a classification accuracy of 90% even when the signal-to-noise ratio was as low as 0.9, ultimately achieving millisecond level recognition of metal/semiconductor CNTs (accuracy of 98%) (Figure 10c). This work further combines the rescan verification mechanism with laser exposure damage control to ensure the reliability of edge equipment analysis. It is worth noting that Zhao et al. [81] focused on scattering dynamics simulations, but their ML method for constructing spin state potential energy surfaces significantly reduced the resource consumption of traditional quantum chemistry (with a 5-fold increase in resolution), revealing the potential of edge intelligence in theoretical chemistry. These studies together prove that edge computing is reshaping the real-time boundary of chemical detection through the ternary collaboration of sensor innovation (flexible electronics/fast spectroscopy), lightweight model (CNN architecture optimization), and physical constraints (material stability/laser control), providing a universal intelligent platform for medical diagnosis, nano electronics manufacturing, and material design.

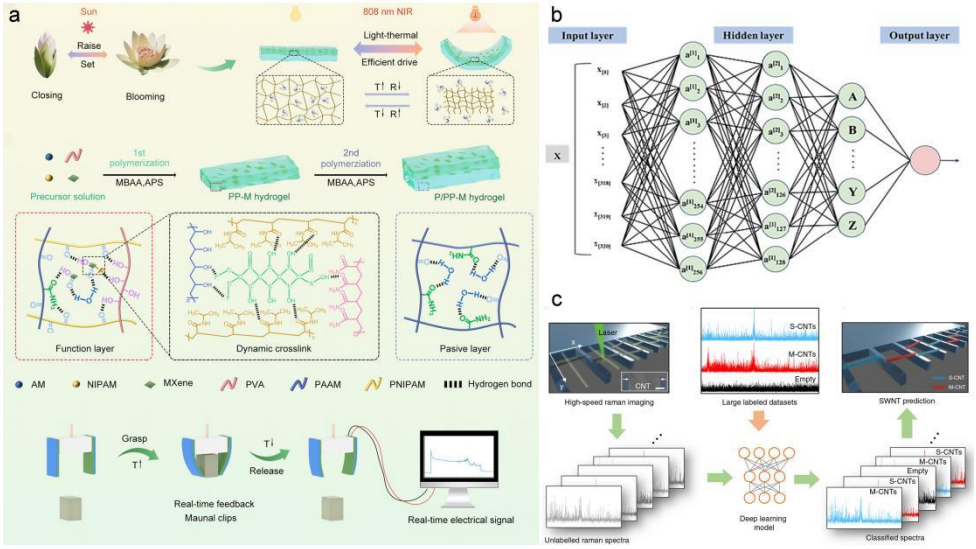

Figure 10. (**a**) Fabrication and light/thermal cascade-driven schematics of P/PP-M actuator[78]. Copyright 2024, Elsevier; (**b**) Schematic illustration of artificial neural network algorithms[79]. Copyright 2024, Elsevier; (**c**) Schematic illustration of deep learning-based Raman spectra analysis for CNT identification[80]. Copyright 2022, Springer Nature.

6.3. Efficient Proxy Model

The deployment of efficient proxy models in resource-constrained systems is significantly accelerating the simulation and optimization of chemical processes. Its core lies in replacing computationally intensive physical models with ML to achieve real-time decision-making at the edge. Hai et al. [82] were the first to apply ANN surrogate models to optimize solid oxide fuel cell (SOFC) systems. By establishing a parameter performance mapping of nanoparticle reinforced electrodes through ANN (such as the influence of compressor pressure ratio and current density on efficiency), traditional thermodynamic simulation is replaced, and multi-objective optimization is driven by a genetic algorithm. Finally, a Pareto optimal solution with an efficiency of 61.8% and a cost of 6.1 $/GJ is achieved in the biogas SOFC supercritical CO_2 mixed system, improving computational efficiency by two orders of magnitude. In the field of nanomaterial design, Jia et al. [83] developed a multi-scale physical simulation ML fusion framework (Figure 11a). Coupling discrete dipole approximation, Monte Carlo and

finite element methods to generate a gold nanorod photothermal dataset, training a lightweight surrogate model to achieve end-to-end prediction of "structural parameters photothermal performance" ($R^2 = 0.972$), reducing spatial temperature field calculation time by 99% and applicable to multi geometric structures such as nanospheres/cages/stars, providing a universal design tool for biomedical photothermal devices. Further expanding to microfluidic simulation, Zhuang et al. [67] proposed a Recurrent Residual Convolution Long Short Term Memory (RRCU ConvLSTM) spatiotemporal prediction architecture, which reduces the prediction error of the electric potential field in the electrowetting process by 40% through boundary constrained loss function, and designs a self cycling mechanism to achieve a single iteration output of 1000 time steps, improving the efficiency of Lattice Boltzmann Electrostatic Simulation by 11 times (Figure 11b). The common innovation of the three lies in the construction of a "lightweight agent model guided by physical mechanisms". These studies collectively demonstrate that efficient proxy models are breaking through the computational bottleneck of traditional numerical simulations by integrating physical prior knowledge (thermodynamic rules/optical transport mechanisms/boundary conditions) with lightweight architectures (ANN/multi-scale ML/spatiotemporal convolution), providing an intelligent computing paradigm deployable on edge devices for energy system optimization, nanomaterial design, and microfluidic control.

6.4. Small Sample Learning

The breakthrough application of small sample learning in chemically embedded systems is significantly alleviating the bottleneck of model generalization in data scarcity scenarios. Its core lies in achieving efficient value mining of limited data through algorithm architecture innovation and physical mechanism guidance. Balraadjsing et al. [84] were the first to construct a QSAR framework based on OECD principles (Figure 11c). Using random forest (RF) and ANN, they successfully predicted the acute toxicity of water fleas to metal nanomaterials under only a hundred in vivo data conditions (AUC > 0.7). Through feature importance analysis, they revealed that molecular structural parameters play a dominant role in toxicity beyond exposure conditions, providing an ethical alternative to animal experimental plans for nanorisk assessment. In the field of real-time sensing, Guo et al. [85] developed a physics algorithm collaborative small sample optimization paradigm. Design a carbon dot fluorescence array with adjustable surface functional groups (-OH/-NH_2), establish a spectral pH quantitative model using GPR (prediction error < 0.5 units), and innovatively use linear discriminant analysis (LDA) to reduce the number of sensors to 2, achieving 100% pH classification accuracy in complex substrates such as milk, significantly reducing the data acquisition burden on edge devices. In response to this, Okeke et al. [86] integrated random forest regression and decision tree algorithm to analyze the temperature-dependent Raman spectra at 80–460 K, addressing the challenge of multivariate coupling in the non-harmonic phonon dynamics of two-dimensional materials (Figure 11d). With only small-scale experimental data, they accurately predicted the redshift and broadening behavior ($R^2 > 0.95$) of the E_2g/A_1g phonon mode in WS_2, and correlated key parameters such as thermal conductivity and thermal expansion coefficient, establishing a universal computational framework for thermal management design of optoelectronic devices. The common breakthrough among the three lies in the complexity of the fusion domain knowledge-constrained model.

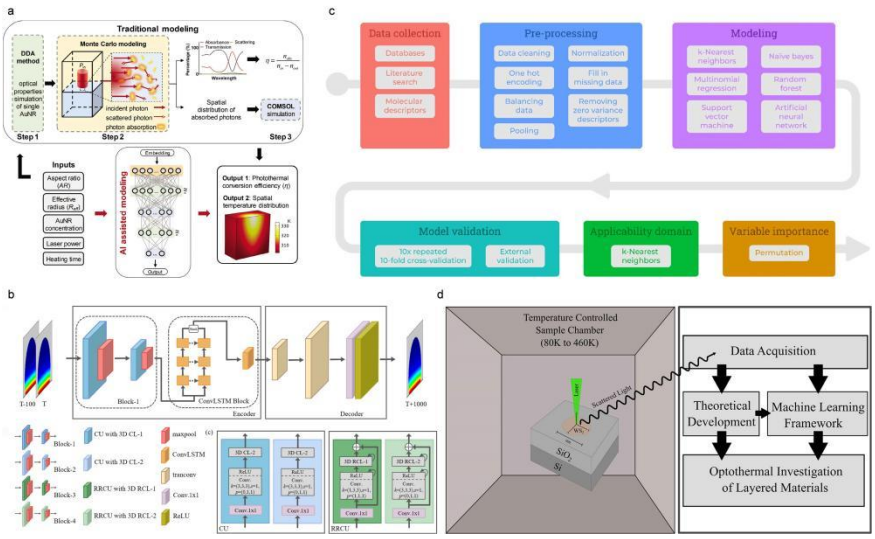

Figure 11. (**a**) Schematic illustration of the photothermal conversion performance prediction of AuNR solution by numerical modeling (black arrows) and AI-assisted modeling (red arrows) [83]. Copyright 2025, Wiley; (**b**) The

proposed network architecture and the different encoder components comprising conventional convolutional units (CUs) or recurrent residual convolutional units (RRCUs) with different k values, where k denotes the size of the 3D convolutional kernel; and the internal structure of CUs and RRCUs [67]. Copyright 2024, American Institute of Physics; (**c**) Diagram of the modeling workflow applied in this study [84]. Copyright 2022, Elsevier; (**d**) Schematic of the experimental setup. Process pathways taken in this work to probe the anharmonic phonons in the bulk WS2 crystal [86]. Copyright 2023, Cell Press.

7. Multimodal Data Fusion

7.1. Cross Database Integration

The breakthrough progress of multimodal data fusion in the field of chemical innovation is significantly reflected in the innovation of cross cross-database integration paradigm. Exner et al. [87] propose a machine operable FAIR metadata management framework to address the long-standing issue of data silos in the field of nanosafety research. This work achieves structured recombination of experimental data through standardized templates (ASINA) and ontology models (DDI Textbook), and designs interdisciplinary interfaces (such as NanoSafety Data Interface) to integrate social science data, providing a semantically aligned multi-source input foundation for ML models. However, its implementation faces core challenges such as the heterogeneity of multidisciplinary data patterns and the disconnection between Electronic lab notebook systems and ML toolchains. These bottlenecks were addressed in the study of He et al., where the NANO. The PTML framework developed by the team innovatively integrates information fusion (IF), perturbation theory (PT), and ML techniques to construct a cross-domain prediction architecture (Figure 12a) [88]. Its high-precision model based on a decision tree (AUROC = 0.97) successfully integrated 440,000 neurodegenerative disease detection data and 260,000 nanotoxicity records, and achieved dual attribute prediction of efficacy/toxicity of neural drug nanocarriers (N_2D_3) in a super high-dimensional heterogeneous data space of 123 drugs, 53 cell lines, and 16 coatings. More importantly, this study validated the engineering value of cross-library integrated data in guiding the optimization of neural drug carriers through experimental computational closed-loop verification, such as Fe_3O_4@CTAB nanoparticle synthesis and 500,000-level virtual screening. The two works form a progressive technical chain. However, the common problems of clinical-level data scarcity and lack of physiological mechanism modeling revealed by both still constrain the generalization ability of the model to clinical translation scenarios.

7.2. Multi-Sensor Collaboration

The multimodal data fusion in the field of chemical sensing is achieving revolutionary breakthroughs through a multi-sensor collaborative architecture, with the core being the decoupling and enhancement of heterogeneous sensing signals by ML. Cruz et al. [89] were the first to construct a carbon-based surface acoustic wave (SH-SAW) sensor array electronic nose system. Complementary response signals were generated by depositing differentiated sensitive layers such as MC/rGO/GO/PDA rGO, and PCA dimensionality reduction and LDA/KNN classification were used to achieve high selectivity recognition of sub-ppm NO_2. Although the system demonstrates portability advantages in multi-gas detection at room temperature, its robustness to complex environmental disturbances such as humidity fluctuations is still insufficient. Shao et al. [90] overcame this limitation in the design of core-shell heterostructures. Their innovative SnO_2@CNS/Pt COFs three-dimensional van der Waals heterostructure maintains resistance to TEA at 95.1@2ppm even under high humidity conditions of 90% by regulating interfacial electron transfer. The PCA-SVM ML framework not only achieves ultra-high response gas classification but also quantitatively analyzes the dynamic calibration effect of core-shell synergistic action on rapid recovery within 5 s. Furthermore, Singh et al. [91] extend the synergistic mechanism to the field of dual gas synchronous detection. Based on the p-n heterojunction superhydrophobic interface formed by MoS_2/SWCNT, differential response trajectories of 0.1 ppm DMF (*p*-type) and 1 ppm NH_3 (n-type) are synchronously captured at room temperature; This study used PCA for trajectory clustering of multidimensional signals, and ML algorithms successfully decoupled the cross sensitivity that is difficult to distinguish in traditional electrical testing, confirming the humidity independent characteristics dominated by charge transfer enhancement.

The common breakthrough lies in using ML to transform the dynamic trajectory of sensing response into programmable structure performance association rules. However, the risk of overfitting in small sample data and the selective balance in complex atmospheres are still scientific bottlenecks that urgently need to be overcome.

7.3. Calculation Experimental Closed Loop

The multimodal data fusion of surface science and materials chemistry is achieving a leap in atomic precision design capability through the computational experimental closed-loop paradigm, with the core of ML bridging the gap between theoretical simulation and experimental observation. Hofmann et al. [92] were the first to establish a dynamic closed-loop verification framework. By using the ML structure search algorithm (SAMPLE) to high-throughput screen the self-assembly configuration space of quinone molecules on Ag (111) surface, combined with DFT energy decomposition to quantify the three force balance mechanism of adsorption site competition, molecular stacking, and steric hindrance effect, accurate matching with STM experiments was achieved with a unit cell size error of <5%, overturning the traditional understanding that "functional group similarity determines assembly configuration". Although this work elucidates the physical laws of static surface assembly, it fails to cover the effects of dynamic variables, such as temperature/coverage on functionality. Laurila et al. [93] overcame this limitation by correlating chemical element properties with cycling. Its pioneering ML deconvolution technique extracts atomic-level chemical elements from X-ray absorption spectroscopy (XAS) of amorphous carbon films, such as sp^2/sp^3 ratios, and constructs regression models to correlate surface functional groups (-COOH/-OH) with dopamine redox rates. Experimental verification shows that carboxylation increases detection sensitivity by two times, providing a programmable surface modification strategy for electrochemical sensors. In addition, Packwood et al. [94] extended the closed-loop paradigm to the functional prediction dimension and developed a DFT-ML embedded simulation framework (Ising model) that accelerated MCMC sampling through the EUF algorithm, accurately predicting the dual functional characteristics of antiferromagnetic order and spin fluctuations of phthalocyanine molecules on the gold (111) surface (Figure 12b).

These all point to the urgent need to develop quantum-accurate MLIP to capture complex processes such as proton transfer and electron correlation.

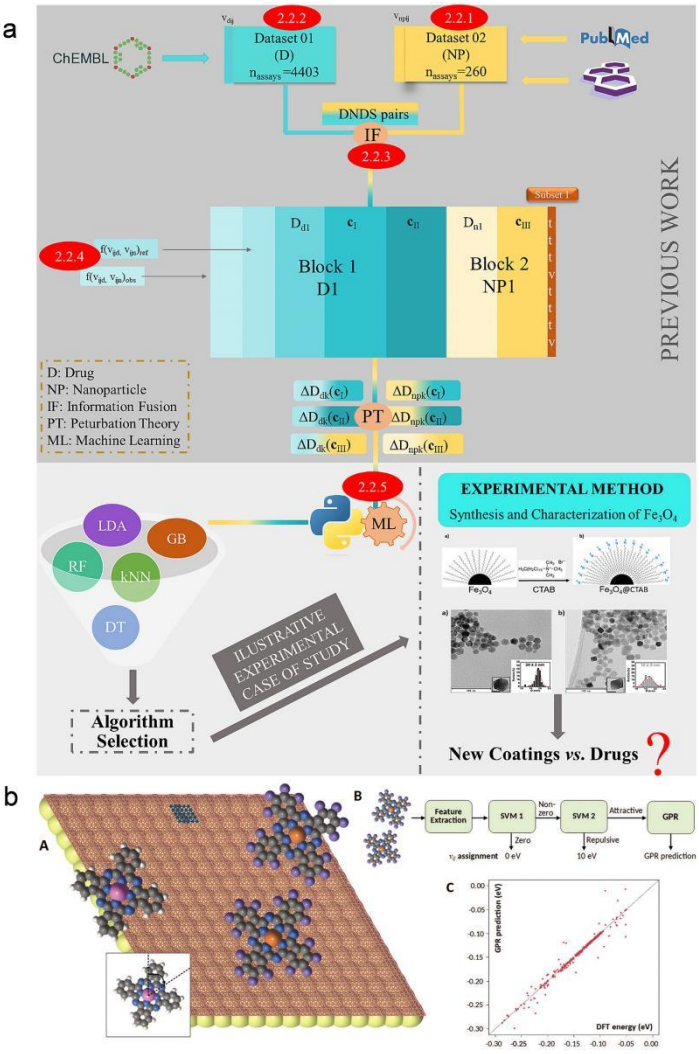

Figure 12. (**a**) IFPTML detailed information-processing workflow [88]. Copyright 2024, BioMed Central; (**b**) Summary of the Ising-like model. (**A**) Illustration of the Ising-like model. This model considers a grid of adsorption

sites (orange mesh), upon which the molecules may adopt one of several orientations (orientation defined in the insert). The adsorption sites available at one gold(111) unit cell are indicated by dark blue. Yellow spheres = gold atoms. (**B**) Protocol for assigning interaction energies to pairs of molecules. The protocol accepts a pair of molecules as input, extracts a feature vector from them, and then applies a series of machine-learned models (SVM1, SVM2, GPR) to assign the interaction energy. See text for details. (**C**) Comparison of interaction energies assigned by the protocol in (**B**) with test data acquired from density functional theory (DFT) calculations, for the case of symmetric molecules. [94]. Copyright 2022, Wiley.

8. Application

8.1. Energy Materials and Catalysis

8.1.1. Battery/Fuel Cell

The efficient operation of SOFC systems urgently needs to solve the multi-scale coupling problem between thermal management and electrochemical performance, and ML is driving a paradigm shift in this field by constructing cross-scale proxy models. Hai et al. [82] were the first to establish a system-level optimization framework. By enhancing heat transfer and electrochemical reaction kinetics through nanomaterials such as electrode functional layer nanoparticles, and using ANN instead of traditional thermodynamic simulations, a parameter performance surrogate model for the biogas SOFC supercritical CO_2 mixing system is constructed; Based on this model, genetic algorithm multi-objective optimization is implemented to determine the optimal operating point in the Pareto balance of efficiency (61.8%) and cost (6.1 $/GJ), achieving a sustainable output of exgo environmental index 0.436. However, the simplification of microscale mass transfer mechanisms in this macroscopic system model limits the adaptability of dynamic loads, especially the unquantified thermal transport operation of nanofluids at the electrode/electrolyte interface. Khan et al.'s mesoscale study addressed this limitation [95]. The ANN model developed by it is based on the Levenberg Marquardt algorithm, which accurately predicts the heat transfer characteristics of graphene oxide/magnesium oxide ethylene glycol nanofluids under the Helmholtz Smolukhov velocity field (error < 0.1%), successfully decoupling multiple slip boundary effects. Quantitatively revealing that thermal slip reduces the temperature gradient by 40% and achieves a 200% increase in heat transfer rate through Joule heating radiation coupling strategy.

In the development of functional materials for advanced battery systems, the dual track collaborative strategy of ML and ligand engineering is becoming a key path to breaking through material performance bottlenecks. The study of vanadium based MOFs confirmed the significant optimization of Zn^{2+} diffusion kinetics by electronic structure regulation through polar ligand engineering (such as–Br group modification) combined with in-situ mechanism analysis (VBr-180 cycling capacity retention rate of 73.53% after 3000 cycles) (Figure 13a) [96]; At the same time, the work innovatively used orthogonal expansion to construct a ML model, successfully decoding the complex nonlinear mapping relationship between electrode coating thickness, current density, and electrochemical performance. Under limited sample conditions, it accurately predicted the working performance and revealed the physical essence of ligand-induced charge redistribution, reducing ion migration energy barriers. This closed-loop research paradigm of "experimental design performance prediction mechanism verification" has been further expanded in manganese-based coordination compound systems [97]. Researchers constructed an extended π–conjugated system (Mn-1,4-DHAQ with a first effect capacity of 138.9 $mAh \cdot g^{-1}$) by precisely controlling the substitution sites of 1,4-dihydroxyanthraquinone (DHAQ), and quantitatively extracted the structure-activity relationship between coating parameters and specific capacity from sparse electrochemical data using a homologous ML architecture (Figure 13b,c). Combined with in-situ UV/XRD multiscale characterization, the core mechanism of para substitution promoting electron delocalization was verified. Two studies jointly highlight the universal value of ML in solving small sample multivariate coupling problems–by establishing a quantitative mapping between electrode manufacturing parameters (thickness/current density) and macroscopic performance, it not only accelerates the screening of high-performance ligands (Br groups in vanadium based MOFs, DHAQ para substitution in manganese based systems), but also reveals the underlying laws of ligand electronic structure regulation of charge transfer efficiency (induction effect enhances redox reversibility, conjugation effect improves electron transfer efficiency), providing a dual drive optimization paradigm for the design of next-generation materials suitable for multi mechanical scenarios of flexible batteries/fuel cells.

The common breakthrough lies in the use of ML to compress high-dimensional physical fields (thermodynamic cycles, electroosmotic flow, radiative transfer) into real-time computable surrogate models. However, they also face the bottleneck of dynamic process modeling and the lack of long-term material evolution

(nanoparticle settling/electrode sintering), which collectively point to the urgent need to develop neural operator networks embedded with physical constraints to unify macro and micro scale prediction.

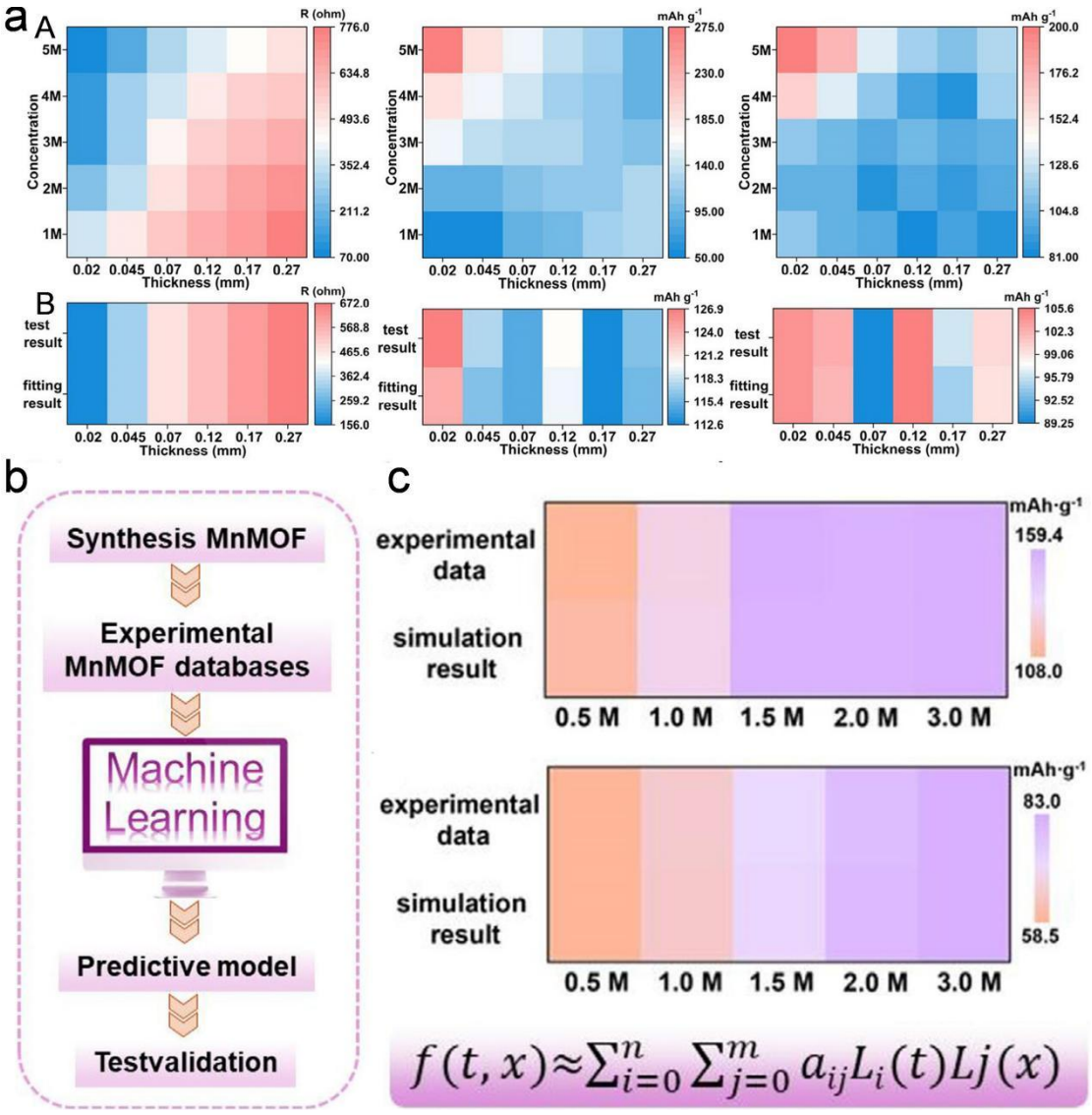

Figure 13. (**a**). **A**) Charge transfer resistance, Specific capacity at small current, Specific capacity at large current. **B**) Machine-learning model training and experimental verification of electrochemical performance in VBr-180 [96]. Copyright 2025, Wiley; (**b**) Workflow adopted to build ML models of MnMOCs [97]. Copyright 2025, Wiley; (**c**) ML model training and experimental verification with current densities of 0.3 $A \cdot g^{-1}$ and 1.0 $A \cdot g^{-1}$, respectively [97]. Copyright 2025, Wiley.

8.1.2. Photothermal Conversion

The rational design of photothermal conversion materials has long been limited by the ultra-high computational cost of multiphysics coupling simulation. Jia et al. [83] constructed a multi-scale numerical framework that integrates the discrete dipole approximation Monte Carlo finite element method to generate a high-dimensional photothermal dataset (1024 efficiency and 2016 temperature fields) (Figure 14a). They trained a ML surrogate model to achieve end-to-end mapping from geometric parameters of nanostructures to photothermal performance (R^2 = 0.972). The computational efficiency has been improved by 99%, breaking through the limitations of traditional simulation structure dependence, and can be extended to heterostructures such as nanospheres/cages/stars, providing a universal tool for intelligent design of biomedical phototherapy devices. At present, it is necessary to overcome the near-field coupling effect of nanoparticle aggregation and the multiple light scattering compensation mechanism of biological tissues.

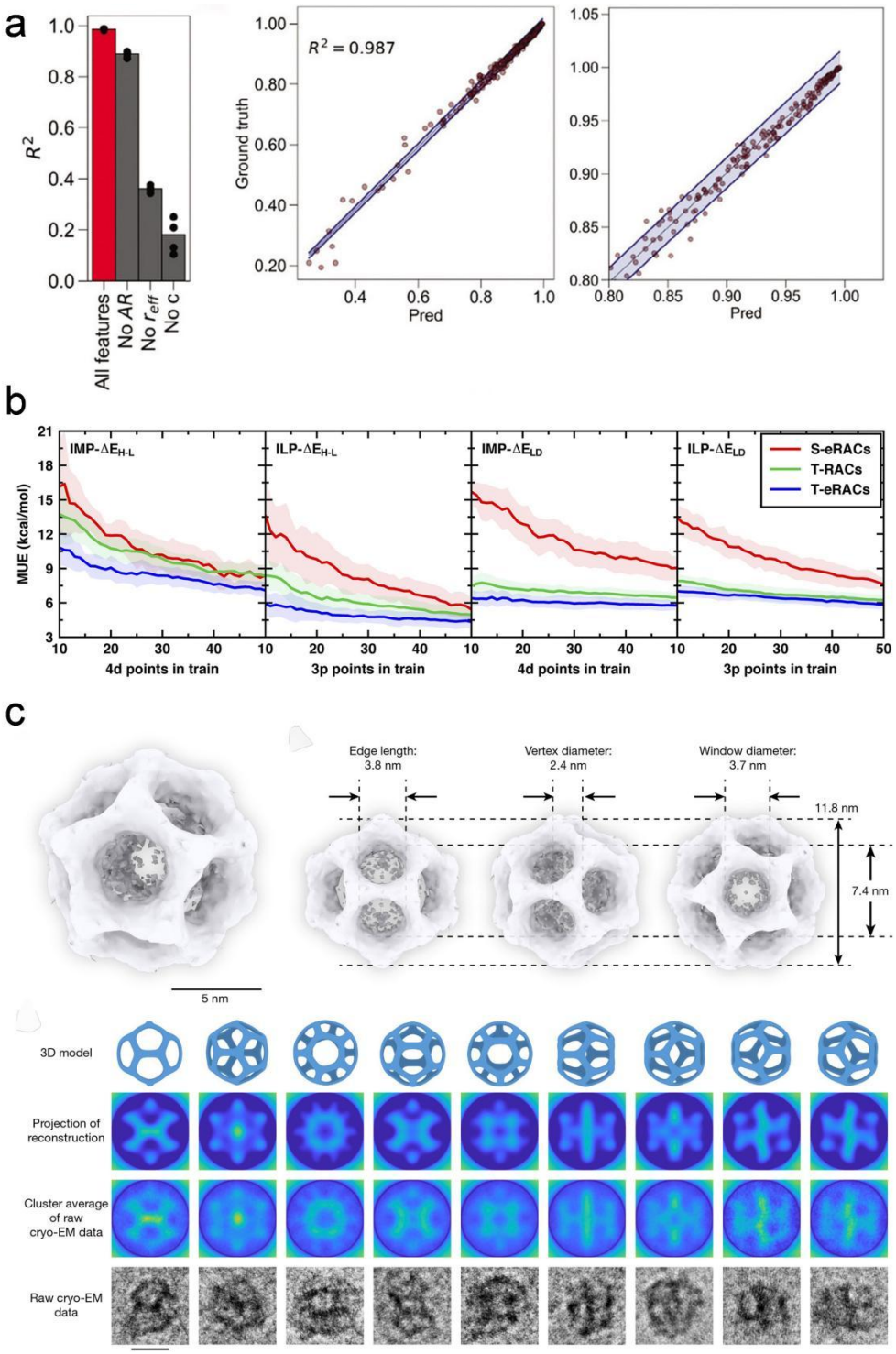

Figure 14. (**a**) Results for photothermal conversion efficiency prediction models [83]. Copyright 2025, Wiley; (**b**) Mean unsigned error (MUE in kcal/mol) for the prediction of ΔEH-L (**left**) and ΔELD (**right**) properties for the isovalent metal pairing (IMP) 3d to 4d learning task and isovalent ligand pairing (ILP) 2p to 3p learning task with addition of training data as indicated on the x axis [62]. Copyright 2022, American Institute of Physics; (**c**) Single-particle reconstruction of the dodecahedral silicage [51]. Copyright 2018, Springer Nature.

8.1.3. Catalyst Design

The rational design of catalysts is achieving a paradigm shift by decoding multi-scale structure-activity relationships through ML, with the core being the collaborative innovation of computational descriptors and experimental structures. Kulik et al. [62] propose a radial basis function descriptor with enhanced electronic properties to address the bottleneck of scarce data for transition metal complexes (Figure 14b). By integrating group number heuristics and nuclear charge information, it accurately encodes the periodic similarity of 3d/4d valence metals. By combining transfer learning frameworks, only about 20 target domain samples are needed to

reduce the prediction error of spin splitting energy by 40%, significantly improving the transferability of open shell complex properties. The breakthrough of this work lies in revealing that eRAC reorders feature space distances to match the periodic law of elements, establishing an efficient computational tool for screening Earth's abundant metal catalysts. However, its descriptors have not yet fully addressed the universality challenges posed by the lack of data on high-spin state complexes and ligand diversity (such as phosphorus ligands vs nitrogen ligands). The complementary breakthrough discovered by Wiesner et al. [51] through mesoscopic structures is beyond this limitation (Figure 14c). By using cryo electron microscopy single particle 3D reconstruction technology, the ultra-small silicon cage (<10 nm) dodecahedral structure synthesized by surfactant micelle directed synthesis was analyzed for the first time, overturning the traditional understanding of the formation mechanism of mesoporous silica. Although the algorithm details were not elaborated in this study, ML-assisted low signal-to-noise ratio image analysis successfully validated the hypothesis of highly symmetrical cage assembly induced by charged micelle templates. The differential modification characteristics of the inner and outer surfaces open up new paths for the design of multiphase catalyst carriers.

The common challenge lies in the need for ML to further integrate quantum chemical constraints and synthetic dynamics: the former requires the extension of eRAC's encoding ability for ligand stereoelectronic effects, while the latter urgently needs to establish a dynamic prediction model for micelle inorganic interface interactions, jointly promoting the paradigm shift of catalyst design from "trial and error synthesis" to "algorithm driven creation".

8.2. Environmental Chemistry and Sustainable Development

8.2.1. Water Treatment Membrane Technology

In the field of environmental chemistry and sustainable development, ML is deeply reconstructing the research and development paradigm of water treatment membrane technology. Ayta ç et al. [98] systematically deconstructed the knowledge graph of 60 years of reverse osmosis membrane research for the first time through the fusion analysis of bibliometrics and natural language processing. Their unsupervised clustering-based sentiment analysis revealed core issues such as membrane fouling control, thin layer composite (TFC) materials, and nano polymer applications, providing a historical coordinate for the development of the field. This macro trend insight directly inspired microscale membrane performance optimization research: Baig's team innovatively used a mixed ML model (ANFIS SVR) to predict the flux of loose nanofiltration membranes (with an accuracy of 95%) in response to the sustainable development needs of dye/salt separation (Figure 15a) [99]. Its feature importance analysis empirically demonstrated the dominant role of operating parameters in separation efficiency, significantly reducing traditional trial and error costs. However, the study revealed insufficient accuracy in predicting retention rates (72%) and challenges in adapting to industrial scenarios, which led to targeted breakthroughs in the high-performance ultrafiltration membrane design by Gao et al. (Figure 15b) [64]. This work pioneered a multivariate prediction framework based on XGBoost/CatBoost, combined with SHAP interpretability analysis to quantify the causal chain of "manufacturing conditions membrane properties performance". It was found that the threshold of nano additives (>1.0 wt%) had a decisive impact on the molecular weight of pore-forming agents, especially elucidating the physical mechanism of high pore-forming agent molecular weight suppressing membrane fouling by reducing pore size. The above data-driven design concept has reached its extreme in the study of stimulus-responsive membranes by Li et al. [100]. The graphene-like material Stimu-C was reverse screened through a generative model, and its MD simulation verified the electric field/pressure synergistic self-cleaning mechanism (Na^+ desorption rate > 95%) and ultra-high water flux (300 L m^{-2} h^{-1} MPa^{-1}), which not only solved the problem of traditional membrane fouling, but also reduced desalination energy consumption to 1/5 of traditional technology.

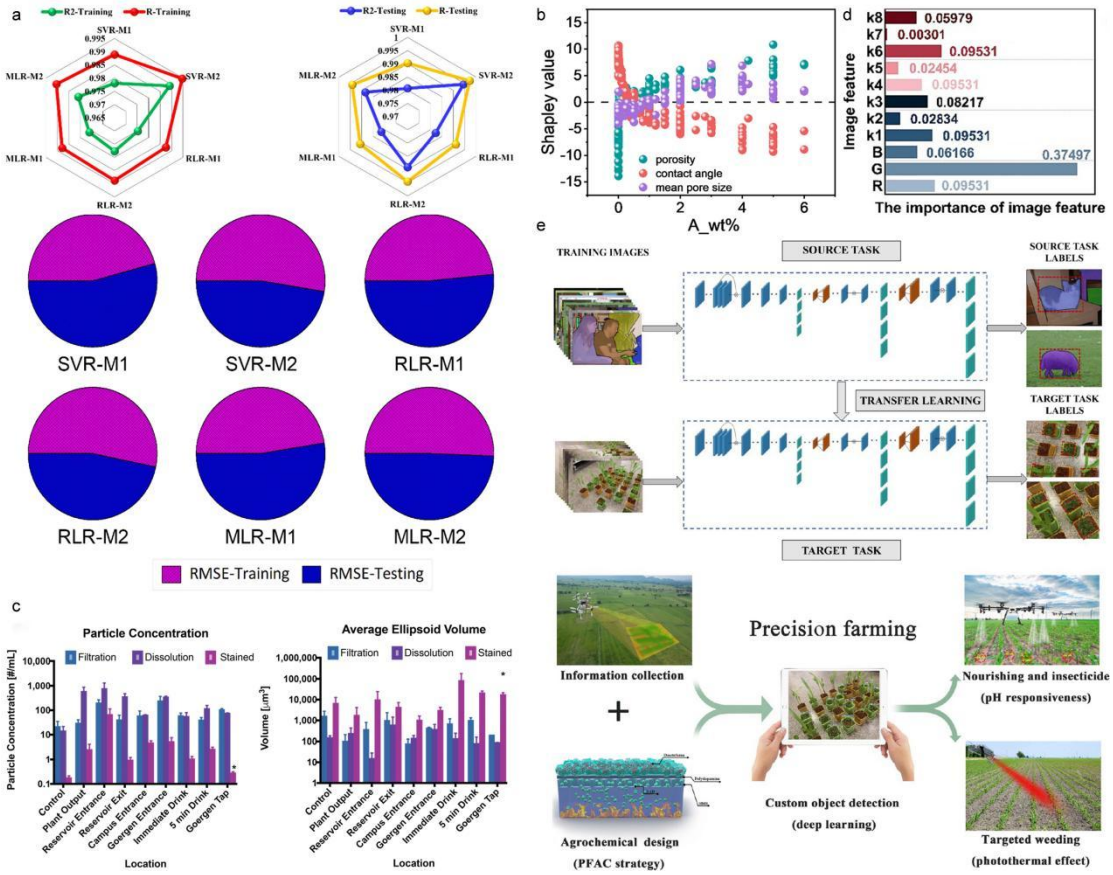

Figure 15. (**a**) Goodness-accuracy and error values for the SVR, RLR, and MLR models for the training and testing phase [99]. Copyright 2023, Elsevier; (**b**) SHAP plot for ML models on membrane properties: Shapley values of A_wt% % for each of the membrane properties [64]. Copyright 2023, American Chemical Society; (**c**) Particle Concentration normalized to the volume of water filtered and average volume of a particle calculated from minor and major axis of image projection [101]. Copyright 2020, MDPI; (**d**) Feature importance retrieved from the Linear model that learns from the training dataset [102]. Copyright 2023, Elsevier; (**e**) The model of the state-of-the-art version of YOLO-v3 is pretrained on the COCO object detection data set and then is trained and tested by transfer learning for a data set including 2822 plant images [103]. Copyright 2021, American Chemical Society.

8.2.2. Pollutant Detection

The pollution detection technology driven by ML is reshaping the accuracy and timeliness boundaries of environmental chemical monitoring. Cruz et al. [89] achieved real-time identification of sub-ppm NO_2 at room temperature for the first time through the collaborative design of carbon-based nano-sensitive layers and SH-SAW sensor arrays, combined with classic ML algorithms such as principal component analysis and linear discriminant analysis (LDA). Their selective enhancement mechanism based on elastic effects provides a low-cost solution for portable electronic noses. This hardware innovation paradigm continues to deepen in the field of gaseous pollutant detection. The three-dimensional van der Waals heterostructure (SnO_2@CNS/Pt COFs) constructed by the Shao team has made a breakthrough in improving the detection response of triethylamine (TEA) to 95.1 (2 ppm) [90], and the core-shell synergistic effect has been analyzed using the PCA-SVM ML framework, maintaining a detection limit of 0.2ppm in 90% high humidity environments, significantly overcoming the environmental adaptability bottleneck of traditional sensors. Similarly, the MoS_2/SWCNT p-n heterojunction sensor developed by Singh et al. [91] utilizes PCA-driven signal decoupling technology to synchronously achieve high selectivity detection of dual gases of 0.1ppm DMF and 1ppm NH_3. ML quantitatively verifies the humidity-independent properties endowed by the superhydrophobic interface, providing a new solution path for complex atmosphere cross-interference problems. In the field of solid pollutant monitoring, Madejski et al. [101] pioneered a silicon nitride nanomembrane filtration ML combined platform, which accurately identifies microplastics < 20 μm through morphological feature quantification algorithms (Figure 15c). The birefringence characteristics verified by its thermal deformation spectrum are correlated with the ML of Nile red staining positive results, revealing the current situation of microplastic pollution in urban water supply pipelines up to 720 particles/mL. However, the interference of impurities (rust/silicate) in small samples still poses a challenge to the robustness of the model.

Regarding liquid heavy metal pollution, Zhang et al. [102] made a breakthrough by directly embedding ML into detection terminals (Figure 15d). A smartphone fluorescence analysis system was constructed based on nitrogen-doped carbon dots synthesized by microwave method in 3 min (90% quantum yield). The innovative *K*-means clustering method was used to extract RGB color space features and combined with ridge regression model, resulting in a detection limit of Cr (VI) in groundwater as low as 0.1574 μg/L and an accuracy rate of >95%. The enhanced selectivity of its internal filtering effect (IFE) and cloud based analysis architecture marked a key step towards civilian and real-time monitoring of pollutants on site.

The current bottlenecks are highly consistent, such as small sample learning, background interference suppression, and real-time optimization of model-embedded deployment, which urgently require the development of adaptive feature engineering and lightweight neural network architectures to cope with the dynamic complexity of environmental media.

8.2.3. Green Synthesis

ML is driving green chemistry towards the goal of atomic economy and zero process waste by reconstructing the paradigm of precise synthesis and intelligent application of nanomaterials. The pesticide fertilizer integrated (PFAC) nanosystem pioneered by Ji et al. [103] achieves near-infrared pH dual response controlled release of herbicides/insecticides/trace elements through a three-layer heterostructure (HMS/PDA/ZIF-8) (Figure 15e). Its core breakthrough lies in the integration of YOLO deep learning models for real-time target detection of weeds in the field, which improves the precision delivery efficiency of herbicides to >90% while reducing the use of agricultural chemicals by 30%; The synergistic mechanism of "photothermal triggering root zone enrichment" revealed in this study (with a 45.9% increase in maize growth) demonstrates the potential of intelligent materials and ML to synergistically optimize resource utilization. However, the recognition robustness and multi-level release dynamics modeling in complex agricultural environments still need to be strengthened. This demand has received methodological support at the level of basic synthesis. Ma et al. [51] used cryo electron microscopy single particle reconstruction technology to analyze for the first time the dodecahedral symmetric structure of <10 nm ultra small silicon cages, and proposed a novel mechanism for surfactant micelle directed self-assembly. The surface of charged micelles guides the oriented arrangement of silica clusters into highly ordered cage-like structures, and their differential modification characteristics on the inner and outer surfaces provide a new platform for pesticide encapsulation.

Two studies jointly point to the core challenge of green synthesis: ML needs to establish interpretable models at both macro-scale applications (dynamic scene recognition in farmland) and micro-scale synthesis (molecular level assembly control) to synergistically achieve a sustainable development loop of "precise design efficient synthesis intelligent application".

8.2.4. Microplastic Monitoring

The microplastic monitoring technology empowered by ML is breaking through the sensitivity and flux bottlenecks of traditional environmental analysis, providing molecular-level insights for tracing plastic pollution. Madejski et al. [101] developed a silicon nitride nanomembrane filtration ML integrated platform, which eliminates background interference through inert substrates and rigorous cleaning, achieving high fidelity capture of microplastics in <20 μm environments for the first time. Combining thermal deformation spectroscopy to verify physical property evidence, and based on ML algorithm to intelligently quantify the morphological characteristics of microplastics in complex matrices (particle/fiber classification efficiency is improved by a hundred times compared to manual methods), the pollution intensity of urban water supply networks can reach up to 720 particles/mL. At present, it is necessary to overcome the problem of insufficient generalization ability for small samples and the ambiguity of <5 μm ultrafine particle morphology. In the future, it is necessary to integrate CNN and a spectral fingerprint library to construct a multimodal recognition framework.

8.3. Biomedical and Nanosafety

8.3.1. Nanotoxicity Prediction

ML is reconstructing the prediction paradigm for the environment and health risks of nanomaterials through multi-scale modeling and interpretability enhancement strategies, but data scarcity and mechanism complexity remain common bottlenecks for cross-disciplinary applications. Balraadjsing et al. [84] were the first to extend the traditional QSAR framework to in vivo endpoint prediction. The random forest model (AUC > 0.7) constructed for acute toxicity of Daphnia magna demonstrated that ML can still identify key toxicological driving factors under

small sample conditions. The contribution weight of molecular structural features to toxicity is significantly higher than that of exposure conditions, providing a feasible approach to reduce animal experiments. Furxhi et al. [65] innovatively developed a Bayesian network (BN) constrained by expert knowledge to address the complexity of environmental media. By integrating probabilistic inference and interpretable rule extraction (such as "particle size < 20 nm and uncoated high toxicity"), it achieved an 82% accuracy in predicting the chronic toxicity of soil silver nanoparticles. Its ability to quantify the synergistic effect of surface treatment and particle size provides a decision threshold for safe and sustainable design (SSbD). However, the incompleteness of ecotoxicological data limits the analysis of deep mechanisms. This limitation was addressed through a breakthrough in the NANO.PTML cross-domain prediction framework proposed by He et al. [88]. This model integrates quantum chemical descriptors of PT, multi-source data integration of IF (700,000 level neurotoxicity nano characteristic dataset), and decision tree ML to achieve an AUROC 0.97 accuracy for predicting the "efficacy toxicity" dual attributes of neural drug carriers. It has been validated through experimental verification of CTAB-coated Fe-O_4 nanoparticles, demonstrating the prospects of computationally guided synthesis. Four studies jointly reveal the core contradiction of nanotoxicity prediction. ML continues to deepen interpretability in feature importance analysis**,** but still struggles to completely overcome the data gap in clinical/environmental real-world scenarios; In the future, it is necessary to develop a multimodal federated learning framework that integrates *in vitro* high-throughput screening, MD simulation, and real-world exposure data to establish a cross level prediction chain from quantum scale interface reactions to ecosystem level transmission.

8.3.2. Biosensing

ML is driving a paradigm shift in molecular recognition from macroscopic concentration monitoring to atomic-level interaction mechanism analysis by reconstructing the signal parsing paradigm of biosensing. The innovative fusion of surface functional group engineering (-OH/-NH_2) and GPR developed by Guo et al. [85] achieves wide range (pH 3–10) high-precision sensing in complex matrices (such as milk) by quantifying the mapping relationship between fluorescence spectra and pH (prediction error < 0.5 units) (Figure 16a); The binary carbon dot optimization strategy established by combining LDA has verified the guiding value of ML for sensor design with 100% classification accuracy, but fluorescence interference suppression in complex environments remains a key bottleneck. This challenge has received a breakthrough response in the field of electrochemical detection. The Hao team pioneered the deep convolutional network FSVNet [104], which extracts 0.2 amol/L copper ion single atom events from high noise backgrounds by analyzing the fast scan voltammetry signals (400 V/s) enhanced by click chemical catalytic amplification and nanomaterials (Figure 16b). The detection limit has been increased by 20 times compared to the traditional signal-to-noise ratio threshold (≥3), achieving single-atom level quantification of metal ions in solution for the first time. The synergistic mechanism of "catalytic amplification deep learning" revealed in this work (with a sensitivity of 4×10^{-19} mol/L) opens up a new path for real-time monitoring of biological catalytic molecules such as CRISPR proteins. However, the delay in click chemistry reaction kinetics restricts real-time applications. The technological evolution of macroscopic ion detection and microscopic atomic recognition mentioned above has formed a methodological convergence at the interface interaction analysis level of biomacromolecules. Stuart et al. [105] constructed a multimodal ML framework that integrates the saturation transfer effect of DISCO NMR with chemical shift fingerprints (such as HPC 4.58 ppm inert proton sites), and can explain the structure-activity relationship of polymer mucin interactions through decision tree modeling (accuracy 0.92). The quantitative correlation between molecular weight (80–150 kDa) and chemical shift not only validates the universality of material-independent descriptors, but also uncovers underestimated proton binding sites, providing mechanism-level guidance for targeted biomaterial design.

Three studies jointly demonstrate the penetrating value of ML in the biosensing chain: from optimizing sensing element design, breaking through physical detection limits, to decoding molecular mechanisms, the algorithm continues to bridge the gap between hardware performance and biological complexity. The current challenges are also highly focused on environmental interference robustness, real-time response capability, and high-dimensional data interpretability. It is urgent to develop adaptive noise cancellation algorithms and lightweight GNNs to achieve cross-scale integration from single-atom events to live dynamic monitoring.

8.3.3. Medical Diagnosis

In the field of medical diagnosis, the accuracy of toxicity prediction of nanomaterials is directly related to the biological safety and clinical applicability of wearable devices. Guo et al. [79] constructed a super-flexible $Ti_3C_2T_x$ MXene/PDMS sponge sensor through hydrogen bonding engineering (Figure 16c). The 92.3% porosity and 9.7 kPa ultra-low modulus characteristics significantly reduced the mechanical irritation of the material, while

the high sensitivity of 14.2 kPa^{-1} and the detection limit of 3.4 Pa achieved super-resolution capture of micro pressure signals, providing a new paradigm for nanoscale biomechanical response monitoring of pathological speech. This work further integrates CNN deep learning architecture and successfully achieves 94% accuracy in classifying speech actions through hierarchical extraction of pressure spatiotemporal features, solving the problem of misclassification of speech features caused by insufficient signal-to-noise ratio in traditional sensors. To complement this technology, Yu et al. [69] developed a superhydrophobic LIG/MWCNT sensor for underwater medical scenarios, achieving a contact angle of 153.4° through SiO_2/Ecoflex interface modification. While suppressing the corrosiveness of body fluids (reducing the risk of nanomaterial leaching), the MLP neural network was used to fuse multimodal strain temporal features, improving the accuracy of Arabic sign language recognition to 99.34%. It is worth noting that both studies have jointly revealed the core challenge of signal stability in nanodevices under environmental interference.

The ML strategies of both (CNN spatiotemporal modeling and MLP feature fusion) have jointly promoted the construction of a closed-loop system for wearable devices from signal acquisition to pathological diagnosis. Its embedded deployment scheme marks a substantial breakthrough in nano-sensing technology towards real-time toxicity monitoring in clinical practice.

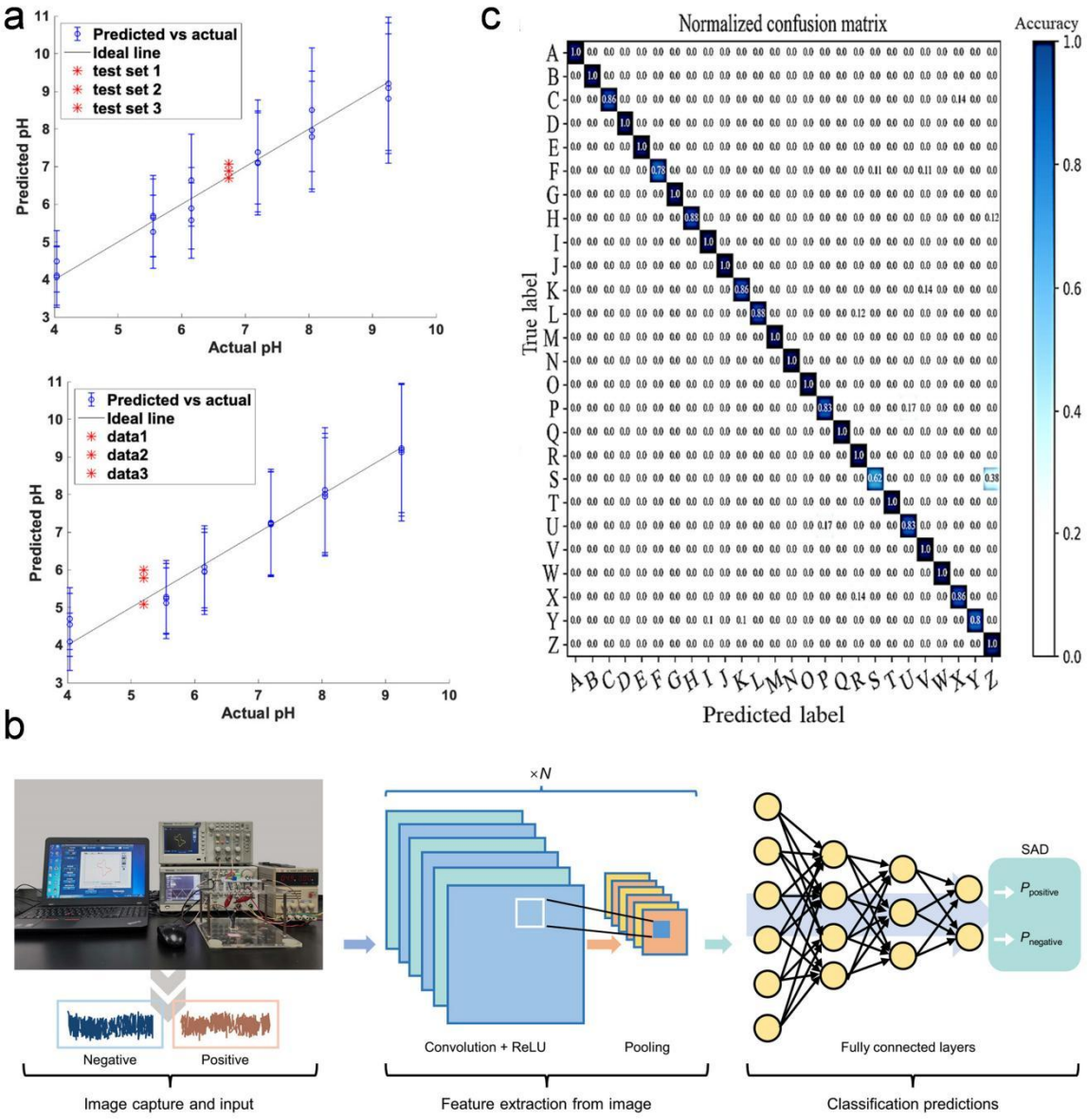

Figure 16. (**a**) GPR prediction by using 2 CDs array built by CD43 and CD69: fresh milk and spoiled milk [85]. Copyright 2024, the Royal Society of Chemistry; (**b**) The framework of our image classifier FSVNet based on DCNN [104]. Copyright 2024, Springer Nature; (**c**) Confusion matrix from ANN for 26 letter pronunciations [79]. Copyright 2024, Elsevier.

8.4. Development of Functional Materials

8.4.1. Intelligent Response Materials

The design paradigm of intelligent responsive materials is undergoing a paradigm transition from experience-oriented to data-driven, with the dual empowerment of ML in structure performance relationship analysis and dynamic behavior optimization being particularly crucial. Inspired by the light response mechanism of water lilies, Jiang et al. [78] constructed a bionic driver with both high strain (1014%) and fast photothermal conversion (808 nm) by using MXene nanosheets to enhance polyacrylamide/PNIPAM double-layer hydrogels. This material system innovatively integrates a strain-sensing unit (GF = 3.62), which converts deformation signals into electrical feedback in real time, forming a "stimulus response perception" closed loop. Its 400 ms response speed and >500 cycle durability provide a reliable execution basis for soft robots. Further research has adopted CNN deep learning architecture to analyze complex deformation patterns and achieve accurate classification of grasping action states, laying the algorithm foundation for solving the core challenge of multi-degree-of-freedom collaborative control. In response to this, Li et al. [100] broke through from the source of material creation by using a ML-driven reverse design strategy to screen out the graphene-like material Stimu-C with a periodic porous structure (pore size 0.5–1.2 nm) from the two-dimensional material configuration space. Its metal-rich properties and 280 GPa Young's modulus ensure mechanical stability. MD simulations reveal the intelligent response nature of the material, where an electric field (1 V/nm) and pressure (50 MPa) synergistically trigger ion desorption kinetics (Na^+ desorption rate > 95%), achieving self-cleaning and anti-fouling functions. The challenges faced by both are also common: the impact of complex environmental disturbances (light fluctuations/organic pollution) on the robustness of material response, as well as the engineering bottleneck of large-scale preparation, which urgently requires the integration of physical constraints in reinforcement learning algorithms for further breakthroughs (Table 3).

Table 3. The physical constraints are incorporated into the reinforcement learning algorithm and the comparison table of strategies.

Material Design	Response Mechanism	ML Strategy	Performance Breakthrough	Common Challenges	Dimension
MXene-enhanced biomimetic hydrogel (strain 1014%)	Light-heat-mechanical energy conversion (closed-loop feedback)	CNN Shape Classification (Control Optimization)	Response time: 400 ms per cycle >500 times	Environmental light interference suppression/Multi-degree-of-freedom control	Thermo-optic responsive actuator [78]
ML reverse screening of Stimu-C porous structure (pore diameter 0.5 nm)	Electrostatic/pressure-assisted ion desorption (self-cleaning)	Generative model + MD verification (structure-activity relationship)	Water flux: 300 $L \cdot cm^{-2} \cdot MPa^{-1} \cdot h^{-1}$/Salt rejection rate: 99.7%	Organic matter pollution simulation/Large-scale preparation	Stimulus-response desalination membrane [100]
ML empowerment contribution	The migration from the bio-inspired to the data-driven design paradigm	Analytical solution of multi-physics field coupling behavior	Cross-scale (macroscopic–microscopic) feature correlation	Precise regulation of the intelligent response threshold and durability boundary	Material-algorithm collaborative adaptability in complex scenarios

8.4.2. Structural Material Reinforcement

In the field of structural material reinforcement, ML is reconstructing the development process of nano-modified composite materials through multi-scale performance mapping and reverse design paradigms. Dong et al. [63] were the first to establish an XGBoost dual objective prediction model ($R^2 > 0.95$), which accurately measured the competitive relationship between compressive strength (UCS) and electrical resistivity (ER) in graphite-based nanocement (GNRCC), and identified key control parameters such as nanofiller ratio and water-cement ratio based on SHAP feature importance analysis. Further, combining NSGA-II multi-objective optimization algorithm to generate a Pareto optimal solution set, providing data-driven design criteria for structural health monitoring materials with synergistic enhancement of mechanical and electrical properties. The prediction optimization framework has been significantly deepened in Tao's work, and his proposed HEStack hybrid integrated architecture (integrating BPNN/RF/XGB) optimizes weight allocation through meta learners, improving the prediction accuracy of compressive strength of nano modified concrete to $R^2 = 0.9924$ (MAPE = 2.84%),

effectively solving the problem of model overfitting caused by nonlinear coupling effects of multiple components such as carbon nanotubes/nano SiO_2. It is worth noting that the interpretability bottleneck of such "black box models" was overcome in the study by Li et al. [56]. Through the dual driving strategy of ANN performance prediction (R^2 > 0.94) and SHAP micro mechanism analysis, the core reinforcement mechanism of nano $Al_2O_3/CaCO_3$ whiskers is revealed, which promotes the phase transition from C-S-H to C-A-S-H (reducing harmful pores by 36.6%) and increases the compressive strength of geothermal concrete at 105 °C by 252%. These findings complement the RReliefF feature weight analysis by Sun et al. [106], whose ANN model (R^2 = 0.885) confirms the dominant role of water cement ratio (contribution weight 35%) in the strength of carbon nanotube reinforced concrete, and accurately captures the nonlinear strength gain law within the range of CNT content 0.05–0.5% (highest + 27.8%).

However, there is still a lack of cross-scale validation and engineering conversion bottlenecks, which urgently require the integration of physically constrained generative models and high-fidelity digital twin technology to break through.

8.4.3. Electronic/Optical Materials

In the development of electronic/optical materials, ML is revolutionizing the development paradigm of functional materials through the improvement of atomic-scale simulation accuracy and reverse engineering of device performance. Della Pia et al. [107] were the first to use an ML potential function (MLP) trained based on first principles to analyze the phase transition behavior of one-dimensional nano-confined water with quantum chemistry accuracy. They determined the precise melting point (280–310 K) of ice nanotubes inside carbon nanotubes (9.5–12.5 Å) for the first time, revealing the negative correlation mechanism between hydrogen bond network attenuation and diffusion coefficient enhancement at the sub-nanometer scale, providing key theoretical support for the thermal management design of nanofluid devices. This breakthrough in atomic scale simulation accuracy was further extended in Mortazavi's work [38], where the developed passive training MTP can reproduce DFT level phonon spectra with only a small amount of ab-initio molecular dynamics (AIMD) data, reducing the computational cost of the thermal expansion coefficient (TEC) of two-dimensional carbon based nanosheets by three orders of magnitude while maintaining high consistency with AIMD (validation at 300–1700 K), significantly accelerating the material screening process of thermosensitive electronic devices. This cross-scale predictive ability from microscopic mechanisms to mesoscopic properties directly empowers the design of macroscopic electronic materials. Vakharia et al. [108] accurately analyzed the structure-activity relationships of 240 metal halide perovskites based on the ElasticNet regression model, not only achieving ultra-low error (MAE = 0.09 eV) in predicting the bandgap of Cs based systems, but also revealing the dominant regulation of halide anions (Cl/Br/I) on the bandgap through DFT-ML joint verification (correlation coefficient 0.98), establishing a lightweight computational tool for optoelectronic device bandgap engineering. It is worth emphasizing that the above basic discovery is a positive driving device-level innovation. Chen et al. [109] utilized the inherent randomness of quantum tunneling current to design a sub-10 nm air channel nanodiode. Its ultra-high entropy true random number generation ability (minimum entropy 0.996 bits/bit) and anti ML attack characteristics (prediction accuracy of 50%) are derived from the precise capture of atomic scale current fluctuations, combined with 80% high extraction rate and 1 ps ultra fast response, marking a transition of encryption hardware primitives from empirical design to quantum effect driven paradigm. The four studies jointly outlined the ML empowerment path of "atomic simulation physical property prediction device optimization". However, the generalization of cross-scale models and industrial manufacturing compatibility remain core bottlenecks that urgently need to be overcome, requiring the integration of multi-fidelity learning and generative design to achieve closed-loop evolution in electronic material development (Figure 17).

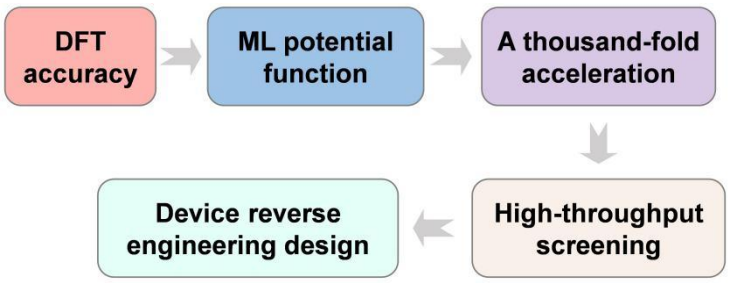

Figure 17. Precision design of ML-enabled path diagram.

8.4.4. Self-Assembling Materials

The design of self-assembled materials is shifting from an empirical rule-based approach to a ML-driven mechanism decoupling paradigm, particularly making breakthrough progress in revealing non-intuitive assembly behavior and multifunctional integration. Jeindl et al. [92] used ML structure search (SAMPLE algorithm) combined with STM experiments to discover for the first time that homologous quinone molecules (B_2O_3/A_2O/P_2O) form three distinct structures (two-dimensional grid/hexagonal ring/parallel row) on Ag (111) surface, overturning the traditional understanding that "the same functional group leads to similar configurations"; The three force balance design principle established quantifies the antagonistic relationship between adsorption site competition, molecular stacking entropy, and steric hindrance effect, revealing that the aspect ratio of the molecular skeleton dominates the assembly path by regulating the relative weights of these forces (with a prediction error of less than 5% per unit cell size), providing a universal theoretical framework for surface functional molecular engineering. This analytical ability for complex interactions extends to the multifunctional design dimension in Packwood's work. The DFT-ML embedded simulation framework developed by it embeds atomic details (DFT) into Ising-like models, and combines "Evolution Under Fire" (EUF) to optimize MCMC sampling, efficiently predicting the dual functional characteristics of antiferromagnetic order and spin fluctuation of phthalocyanine molecules on Au (111) surface [94]. It proves that asymmetric ligands can induce both structural order and magnetic fluctuation, creating a new type of noise source for spintronic devices. It is worth noting that the breakthroughs in the surface assembly mechanism mentioned above are driving the transformation of solution-phase self-assembly. Ma et al. [51] used cryo electron microscopy single particle 3D reconstruction technology to analyze the dodecahedral structure of <10 nm silicon cages and found a new mechanism for the directional self-assembly of silica clusters guided by charged micelle surfaces. The potential for differential modification of the inner and outer surfaces (although the ML algorithm is not described in detail) demonstrates the paradigm shift of template engineering from passive observation to active design. Three studies jointly outline a three-level path for ML to crack the complexity of self-assembly. It is urgent to develop physically constrained reinforcement learning frameworks that integrate real-time environmental variables to promote self-assembly from static structural design to dynamic functional programming (Figure 18).

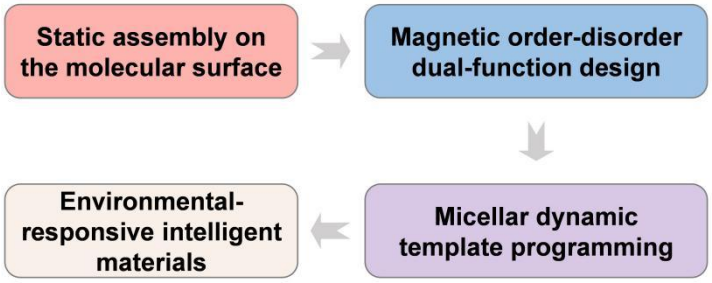

Figure 18. The reinforcement learning framework integrates the real-time environmental variable materials design path diagram.

8.5. Innovation of Basic Chemistry Research Tools

8.5.1. Automated Characterization

The intelligent revolution of representation tools is reconstructing the observational dimensions of chemical research through ML, achieving a paradigm shift from static analysis to dynamic closed-loop (Table 4). Boiko et al.[75] were the first to develop a modular AI pipeline (classical CV and deep learning), which achieved real-time electron microscopy analysis of the dynamic process of ionic liquid phase for the first time. Its denoising segmentation tracking process not only reveals the anisotropic regulation mechanism of electron beam scanning mode on droplet motion, but also opens up a universal framework to promote cross-domain adaptation. This ability to capture dynamic processes has been extended to the macroscopic scale in Zhang's work [80]. By combining Raman line scanning strategy with CNN classifier, the recognition speed of suspended carbon nanotubes is increased by 50 times. Softmax threshold optimization achieves a metal/semiconductor classification accuracy of 98% (signal-to-noise ratio of 0.9), and millisecond-level exposure synchronously suppresses laser thermal damage, establishing a new online quality inspection standard for nanoelectronic device manufacturing. In the dimension of microstructure analysis, Kho et al. [73] systematically verified the universality of the unsupervised clustering workflow. UMAP integrates 4D-STEM crystal orientation and EDX composition data (F1 score improved by

22%), which is significantly better than VAE which requires fine-tuning. Its proposed unsupervised evaluation index for intra-cluster density/inter-cluster separation achieves automated segmentation of microstructures in multiple systems such as $Co_2FeSi/AuAl_2Cu$, breaking the efficiency bottleneck of traditional manual labeling. It is worth emphasizing that the aforementioned technological advancements are driving the formation of a closed-loop system for instrument autonomous decision-making. Wang et al. [110] integrated AdaBoost spectral classifier (98.5% surface state recognition accuracy) and a watershed terrain analysis algorithm to construct an STM cutting-edge "acquisition evaluation electric shock adjustment" fully automatic control system. Through data augmentation, the problem of small sample learning for high noise *dI*/*dV* spectra was solved, saving 90% of instrument operation time. These findings echo the methodology of Guccione's innovation [111], which recursively quantifies feature descriptors to transform atomic pair distribution functions into interpretable temporal features. Combined with KNN crystal system classification (balance accuracy of 81%) and vector superposition algorithm, it achieves a matching rate of 90% for the top 10 monoclinic cell parameters even under 40% crystal impurity interference, opening up a new path for structural analysis of disordered materials that bypasses diffraction indices. Ultimately, this will drive chemical research from "manual observation, empirical judgment" to a new era of "machine perception, algorithmic decision-making". The core that urgently needs to be broken through lies in the standardization of cross-platform protocols and the interpretability of physical mechanisms. It is necessary to develop a federated learning framework embedded with first principles to integrate multi-source representation data and achieve a closed-loop leap in chemical cognition (Table 5).

Table 4. Matrix Analysis of Automated Characterization Techniques.

Technical Level	Key Breakthrough	ML Empowers Value
Dynamic process analysis	Real-time tracking by liquid phase electron microscopy [75] Raman line scanning for high-speed classification [80]	Convert transient phenomena into quantifiable time-series data
Multimodal data fusion	UMAP integrates 4D-STEM + EDX [73] PDF Recursive Feature Engineering [111]	Break through the limitations of a single representation method in information acquisition
Instrument autonomous decision-making	STM advanced closed-loop regulation [110] Softmax threshold optimization [80]	Replace manual operation with 24/7 stable characterization
Adaptability to extreme conditions	Improvement of PDF parameters under 40% noise level [111] Low Signal-to-Noise Ratio CNT Classification [80]	Low Signal-to-Noise Ratio CNT Classification
Common Challenges	Physical mechanism black box [75] Missing cross-platform interface [110] High-order feature interpretability [73]	Calling for physical embedding algorithms and standardized protocols

Table 5. Human-machine collaboration reconfiguration.

Traditional Model	Intelligent Mode	Efficiency Improvement
Artificial electron microscopy observation	Real-time droplet tracking by AI	The processing efficiency has been increased by 100 times
Advanced adjustment of the STM experience	Closed-loop automatic control	The operation time has been reduced by 90%.
Diffraction Expert Index	PDF ML Analysis	The noise tolerance has been increased by 40%

8.5.2. Computational Chemistry Acceleration

Machine learning potential function (MLIP) is used to reconstruct the simulated boundaries of computational chemistry, breaking through the efficiency bottleneck of traditional force fields and first principles by integrating quantum accuracy and mesoscopic scale (Figure 19). Ghorbani et al. [58] constructed an MLIP for BeN_4/MgN_4/PtN_4 two-dimensional materials based on MTP, and for the first time revealed the anisotropy of thermal conductivity and Young's modulus with DFT accuracy. The directional performance of the armchair is significantly better than that of the serrated direction (such as BeN_4 having a thermal conductivity 1.8 times higher), and the anisotropy of PtN_4 decays faster with increasing temperature, providing key design parameters for nano thermal management devices; This method reduces the computational time of MD by two orders of magnitude, but the high temperature lattice vibration nonlinearity still challenges the model's generalization ability. This precision efficiency balance paradigm was extended to extreme condition reaction simulations in Lindsey's work [59], where the developed ChIMES MLIP achieved high-pressure kinetic calculations of millions of atom-level C/O systems. It was found that carbon clusters undergo "atomic free riding" (carbon atoms migrate through chemical bonds) rather than the classical diffusion Ostwald ripening mechanism, providing a scalable silicon-based design tool for the synthesis of nanocarbon materials. It is worth noting that the computational acceleration requirements of complex particle systems have given rise to more universal solutions. Isfeldt designed a geometric invariant neural network that processes rigid body coordinates through a rotation/translation invariant abstraction layer to directly predict the force/torque of carbon nanotubes and other particles (with an error of <0.1%) [60]. This method is 100 times more efficient than traditional coarse-grained methods and has 12.5% noise robustness. However, its generalization in multi-dispersed systems urgently needs to be verified. In the field of mesoscopic fluids, Zhuang et al. [67] proposed the Deep Spatio Temporal Prediction Architecture (RRCU ConvLSTM), which optimizes the prediction of the potential field with boundary constrained loss function (reducing the error by 40%), achieving an 11 fold acceleration of lattice Boltzmann simulation during the electrowetting process; Its self cycling mechanism predicts 1000 time steps in a single iteration, but the relaxation time sensitivity reveals the deep requirements for embedding physical constraints. The core that urgently needs to be broken through lies in the fidelity of cross-scale transmission and the algorithm embedding of physical constraints. It is necessary to develop a multi-fidelity active learning framework to adaptively allocate computing resources and achieve a new era of computational chemistry from "efficiency compromise" to "precision controllability".

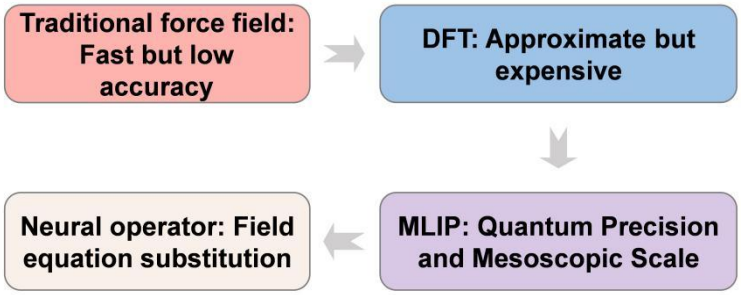

Figure 19. A schematic diagram illustrating the path of breaking through the performance bottlenecks of traditional force fields and first-principles methods through the integration of quantum precision and mesoscopic scale.

8.5.3. Literature Knowledge Mining

Literature knowledge mining is reconstructing the cognitive paradigm in the field of chemistry through ML-driven multimodal analysis, shifting from empirical induction to data-driven domain knowledge graph construction (Figure 20). Ayta ç et al. [98] integrated bibliometrics and ML (NLP clustering/sentiment analysis), based on 1424 reverse osmosis membrane literature, quantitatively revealed for the first time the dual theme competition evolution law of "membrane composite material modification" and "membrane fouling control". By using LDA topic clustering recognition technology to identify subdomains, the sentiment model tracks the transition of domain discourse from technology-oriented (1960s) to application-neutral (2010s), and quantifies the positive correlation between technology maturity and objectivity ratings. At present, it is necessary to overcome three major challenges: historical data heterogeneity, dynamic modeling of topic mutations, and ambiguity in technical text sentiment annotation. It is urgent to develop cross-era semantic alignment algorithms to empower reverse design of membrane materials.

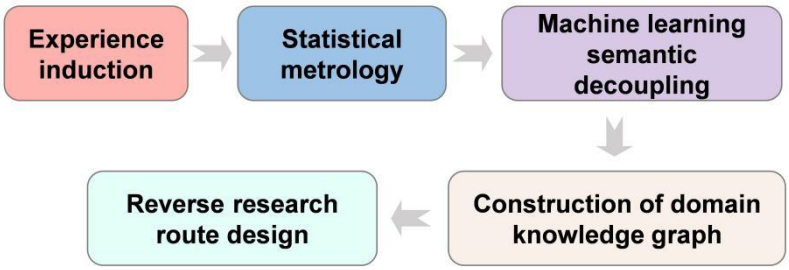

Figure 20. A schematic diagram illustrating the path of transforming from empirical induction to data-driven construction of domain knowledge graphs through ML.

8.5.4. Standardized Database

In the process of promoting the transformation of the chemical research paradigm towards data-driven, the construction of standardized databases is the core link of fundamental tool innovation, and its core challenge lies in overcoming the interoperability and machine-readable barriers of multi-source heterogeneous data. Exner et al. [87] proposed a metadata governance framework based on real-time FAIR principles, which achieves cross-platform semantic interoperability through ASINA standardized templates and DDI ontology models, and integrates electronic experiment notebooks (ELNs) to achieve automatic structured capture of experimental data, completely solving the problem of chemical data silos. This framework provides high-quality structured input for ML and supports risk prediction in complex fields such as nanosafety. However, it needs to overcome three major bottlenecks: differences in interdisciplinary data patterns, ELN-ML integration gaps, and a lack of incentive mechanisms for data sharing.

8.6. Industrial and Agricultural Intelligent Systems

8.6.1. Precision Agriculture

Under the framework of industrial and agricultural intelligent systems, the core challenge of precision agriculture is to achieve efficient, targeted delivery and dynamic regulation of agricultural chemicals, in order to break through the ecological risks and resource waste bottlenecks caused by traditional extensive pesticide application. Ji et al. [103] developed a nano deep learning collaborative pesticide fertilizer integrated (PFAC) system: a three-layer nanocarrier (HMS loaded herbicide/PDA photothermal layer/ZIF-8 loaded insecticide) achieved spatiotemporal controllable release through NIR-PH dual response; Integrating YOLO model for real-time analysis of farmland images, driving near-infrared devices to accurately activate weed control (field weed control rate >90%), synchronously reducing pesticide usage by 30% and increasing corn root length by 45.9%. This closed-loop intelligent system achieves full chain optimization from material design to field decision-making.

However, the integrated system still faces key challenges: dynamic disturbances in complex field environments (such as variable lighting and crop occlusion) continue to constrain the upper limit of target recognition accuracy, requiring the development of lightweight and anti-interference model architectures that are more suitable for agricultural scenarios. In addition, the precise modeling of multi-level response release kinetics within nanocarriers (involving NIR penetration depth, pH microenvironment feedback, and diffusion rate coupling) has not fully integrated ML methods (such as physical information-based neural networks), which limits the system's predictive control ability in variable environments.

8.6.2. Nano Lubricants

Under the framework of industrial intelligent systems, the development of nanolubricants urgently needs to break through the traditional trial-and-error research and development mode, and achieve predictable design and precise control of tribological properties. He et al. [112] provide innovative solutions for the field of intelligent lubrication by integrating novel nanomaterial designs with ML prediction models (Figure 21a). This study pioneered the use of graphene acetylene as a functional additive in water-based nanofluids. Experimental characterization (needle disk friction test) confirmed that it can significantly optimize the friction interface behavior at extremely low concentrations (0.3 wt%). The friction coefficient decreased by 18%, the wear rate decreased by 49.5%, and a composite friction protective film with a thickness of about 6.4 nm was synergistically formed.

8.6.3. Concrete Optimization

Under the framework of industrial intelligent systems, the performance optimization of concrete materials is undergoing a paradigm shift from experience-oriented to data-driven. The core challenge lies in deciphering the multi-scale coupling mechanism between nano modifiers, cement matrix, and environmental factors, and achieving a predictable design of complex formulations. Li et al. [56] first proposed the "interpretable ML micro characterization" dual drive strategy (Figure 21b). By using ANN to predict the performance of geothermal tunnel slag concrete with high accuracy ($R^2 > 0.94$), and using SHAP algorithm to analyze the decisive influence of nano Al_2O_3 and $CaCO_3$ whiskers on pore structure (such as 0.5% Al_2O_3 reducing harmful pores by 36.6%), industrial grade formula design can be guided (1% whiskers + 0.5% Al_2O_3 increasing compressive strength at 105 °C by 252%); Microscopic analysis further revealed that its enhancement essence is the C-S-H to C-A-S-H phase transition induced by nano fillers and pore refinement (diameter reduction of 104.8 nm), providing a physical and chemical basis for ML models. The paradigm of "mechanism-inspired intelligent design" has been extended to the dimension of ultra-high-precision prediction in the work of Tao et al. For multi-nano component (SiO_2/CNT, etc.) concrete systems, the HEStack hybrid integrated architecture was developed [57]. By stacking BPNN, RF, XGB based models and optimizing the weights of meta learners, the breakthrough accuracy of compressive strength prediction was achieved on 94 groups of small sample data ($R^2 = 0.9924$, MAPE = 2.84%), and its anti overfitting characteristics were significantly better than that of a single model (such as BPNN-MAPE > 10%), which established a new standard for industrial high-throughput formula screening. Sun et al. [106] focus on feature engineering and causal analysis (Figure 21c). Based on 282 sets of carbon nanotube (CNT) concrete data, the superiority of the ANN model in capturing nonlinear reinforcement effects was verified ($R^2 = 0.885$ vs. KNN 0.838), and the RReliefF algorithm was used to quantify the contribution of key factors (water cement ratio of 35%, cement content of 28%), accurately characterizing the strength gain law within the range of CNT dosage of 0.05-0.5% (up to 27.8%), providing decision-making basis for low-cost formula optimization.

Collaboration and challenges from the perspective of ML: Three studies jointly demonstrate the deep application of ML in the concrete industry, from the interpretability optimization of ANN, the improvement of generalization ability of ensemble learning, to causal inference of feature importance, gradually constructing a "mechanism prediction regulation" closed loop. However, the field still faces common bottlenecks:

1. Data limitations: Long-term performance data (such as geothermal aging and creep) and extreme operating conditions (high temperature/high pressure/corrosion) samples are scarce;

2. Model extrapolation risk: The physical constraints of nanoparticle dispersion stability, multi-component interaction effects, and interface dynamics have not been fully integrated.

3. Lack of cross-scale correlation: There is a lack of a mathematical expression bridge between micro mechanisms and macro performance predictions. In the future, it is necessary to develop ML models embedded in physics, combined with in-situ monitoring data streams, to achieve intelligent control of the entire life cycle of concrete.

8.6.4. Hardware Security

In the security architecture of industrial intelligent systems, hardware-level random number generators are the core foundation for building cryptographic primitives, and their entropy source quality directly determines the anti-cracking ability of the encryption system. Chen et al. [109] achieved high entropy true random number generation based on intrinsic fluctuations of quantum tunneling current through innovative design of sub-10-nanometer air channel nanodiodes, providing a revolutionary solution for the next generation of secure hardware.

This work marks a paradigm shift in hardware security from "algorithmic defense" to "physical immunity". Its quantum tunneling entropy source mechanism provides a foundational security element that can resist ML attacks for highly sensitive scenarios such as blockchain and quantum communication, fundamentally strengthening the trust chain of industrial intelligence systems.

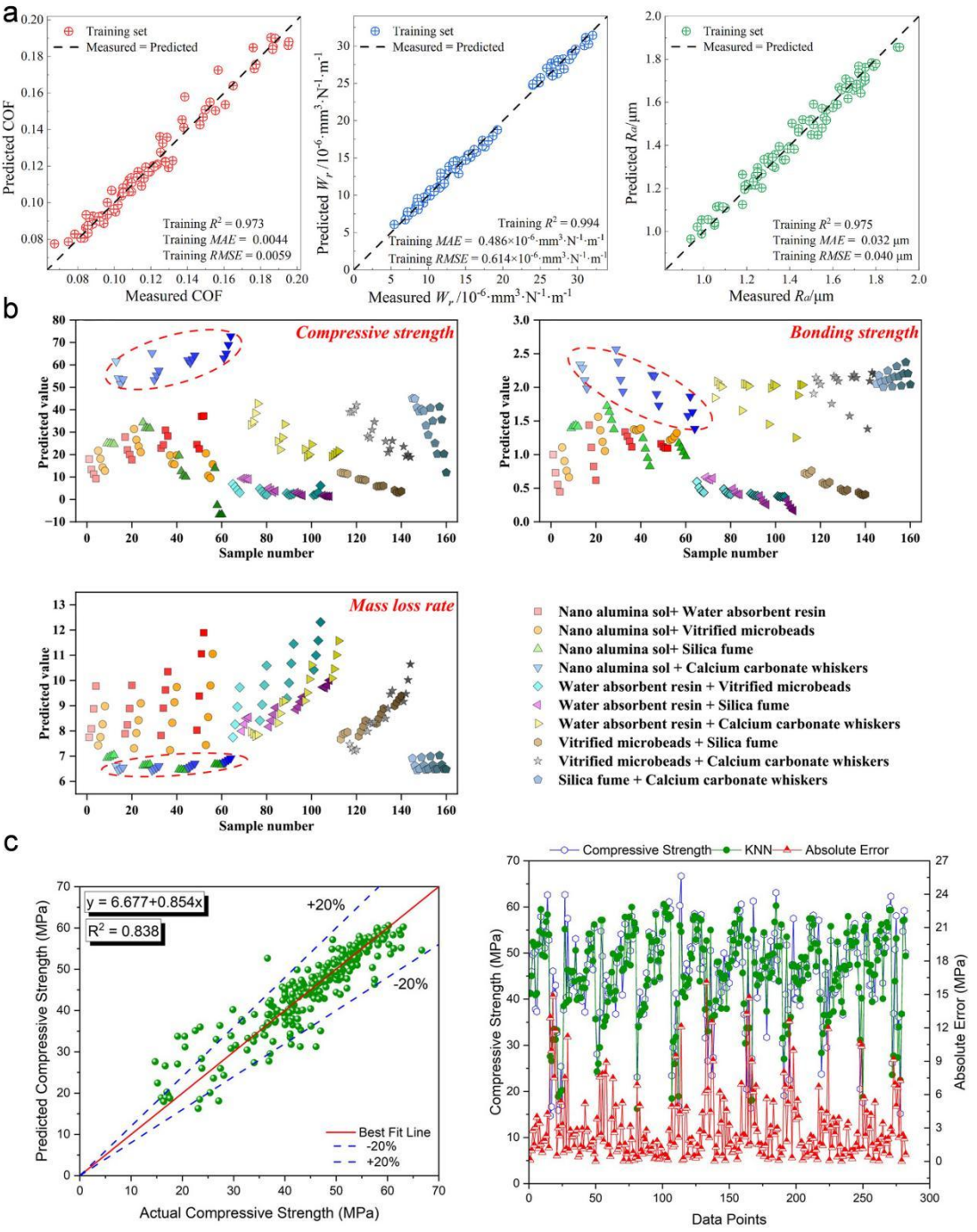

Figure 21. (**a**) Outcome predictions of the RF algorithm applied to the training dataset categorized by distinct output metrics (COF, Wr, and Ra) [112]. Copyright 2025, Elsevier; (**b**) The compressive strength, bonding strength, and mass loss rate of modified FTMS [56]. Copyright 2025, Elsevier; (**c**) KNN model outcomes R^2 graph and error propagation graph [106]. Copyright 2025, Elsevier.

9. Summary and Challenge

The application of ML in the field of chemical innovation has made four key advances:

(1) Generative design framework: effectively addresses combinatorial explosion problems, such as diffusion models (such as PXRDnet) that extend crystal structure prediction to 7 types of systems with an accuracy rate of 80%.

(2) Geometric deep learning: Based on SE (3)–equivariant graph neural network, a force field with quantum chemical accuracy (error < 1 kcal/mol) was established, effectively solving the conformation dependence problem in chiral recognition.

(3) Explainable ML and artificial intelligence collaborative design: Combined with SHAP analysis and working condition characterization, the C-S-H to C-A-S-H phase transition induced by nano fillers was verified by significantly reducing the micropore rate.

(4) Edge intelligent systems have promoted the development of autonomous laboratories, specifically reflected in the scanning electron microscope defect synthesis driven by c-GAN (95.31% accuracy) and solar seawater desalination controlled by Online Sequential Extreme Learning Machines (OS-ELM) (61.14% efficiency).

Despite significant progress, the following core bottlenecks still require interdisciplinary collaboration to overcome:

(1) Physical mapping limitations: Rigid coordinate mapping methods (such as CMED) have limited applicability in low symmetry systems, and there is an urgent need to develop SE (3)–adaptive manifold methods that integrate density functional constraints. The latest developments in the field of geometric deep learning, especially the SE (3) equivariant converter architecture, demonstrate the prospect of developing more flexible and adaptable coordinate representations that can handle low-symmetry systems through learnable attention mechanisms.

(2) Obstacle to dynamic process modeling: Existing MLIP (such as MTP/ChIMES) are difficult to effectively describe non-equilibrium dynamic processes and need to be improved by combining neural operators with stochastic thermodynamic theory. Combining neural operator networks (such as Fourier neural operators) with multi-body potential energy frameworks provides a promising direction for capturing multi-scale dynamic changes while maintaining computational efficiency and physical consistency.

(3) Data scarcity cycle: The severe scarcity of high-fidelity data in the three-slope system severely limits the training and application of ML models. It is recommended to establish a cross-institutional blockchain database that follows the FAIR principles (discoverable, accessible, interoperable, and reusable). The federated learning architecture provides a feasible solution for utilizing distributed data sources while protecting privacy, enabling institutions to collaboratively train models without centralizing sensitive data. In addition, generative models and transfer learning techniques can help alleviate data scarcity issues through intelligent data augmentation and knowledge transfer from related fields.

(4) Interpretability illusion: Although interpretable artificial intelligence (XAI) tools like SHAP and symbolic regression are very useful for providing insights, they can also be misleading. On small or biased datasets, SHAP may assign false correlations to high importance, thereby "explaining" noise rather than the true underlying physical principles. Similarly, symbolic regression may generate overfitting or physically unreasonable equations. There is an urgent need to develop XAI methods that can cope with data limitations and are based on physical constraints to avoid erroneous mechanistic interpretations. Integrating physics-based regularization methods and causal discovery frameworks into XAI methods can help distinguish between false correlations and physically meaningful relationships. The Bayesian method can provide uncertainty quantification, thereby achieving more reliable explanations.

(5) Embedded deployment challenges: Nanosensor drift (such as MXene/PDMS interface) affects the reliability of edge computing systems, so it is necessary to develop and implement a digital twin system based on physical constraints for optimization. The advanced digital twin framework combines online learning capabilities with physically aware calibration algorithms, which can continuously adapt to sensor drift and environmental changes, and maintain system reliability through self-correction mechanisms and transfer learning methods.

The future development direction lies in building an intelligent chemistry research paradigm, which integrates a federated learning architecture based on first principles rules with an automated robot platform to achieve precise synthesis of atomic precision programmable substances.

Author Contributions

H.Z.: conceptualization, methodology, software, writing—reviewing and editing, writing—reviewing and editing; W.Z.: writing—reviewing and editing, investigation; S.G.: data curation; J.X.: data curation; J.Z.: visualization; M.S.: writing—reviewing and editing; J.X.: investigation; L.L.: validation; H.P.: supervision, writing—reviewing and editing; J.D: supervision, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (NSFC-22509178, 52371240), China Postdoctoral Science Foundation (2025M770225), Yangzhou Innovation Capability Enhancement Program (YZ2022170).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

We declare that we have no financial and personal relationship with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any data, service, product and/or company that could be construed as influencing the position presented in the manuscript entitled.

Conflicts of Interest

The authors declare no conflict of interest. Given the role as Editor in Chief, Prof. Huan Pang had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-assisted Technologies

No AI tools were utilized for this paper.

References

1. Pihlajamäki, A.; Matus, M.F.; Malola, S.; et al. GraphBNC: Machine Learning-Aided Prediction of Interactions Between Metal Nanoclusters and Blood Proteins. *Adv. Mater.* **2024**, *36*, 2407046. https://doi.org/10.1002/adma.202407046.
2. Li, Q.; Zhang, Y.; Guo, X.; et al. Nucleation and Growth Mechanisms of Micro/Nano Structural Manganese-Trimesic Acid Coordinations for Aqueous Zinc-Ion Batteries. *Angew. Chem. Int. Ed.* **2025**, *64*, e202509741. https://doi.org/10.1002/anie.202509741.
3. Zhang, G.; Lu, Y.; Yang, Y.; et al. Dynamic Phase Transformations of Prussian Blue Analogue Crystals in Hydrotherms. *J. Am. Chem. Soc.* **2024**, *146*, 16659–16669. https://doi.org/10.1021/jacs.4c03827.
4. Zhang, G.; Feng, W.; Du, G.; et al. Thermodynamically-Driven Phase Engineering and Reconstruction Deduction of Medium-Entropy Prussian Blue Analogue Nanocrystals. *Adv. Mater.* **2025**, 2503814. https://doi.org/10.1002/adma.202503814.
5. Dahl, J.C.; Niblett, S.; Cho, Y.; et al. Scientific Machine Learning of 2D Perovskite Nanosheet Formation. *J. Am. Chem. Soc.* **2023**, *145*, 23076–23087. https://doi.org/10.1021/jacs.3c05984.
6. Chen, L.; Yang, S.; Li, Y.; et al. Precursor Symmetry Triggered Modulation of Fluorescence Quantum Yield in Graphene Quantum Dots. *Adv. Funct. Mater.* **2024**, *34*, 2401246. https://doi.org/10.1002/adfm.202401246.
7. Le, T.C.; Yan, B.; Winkler, D.A. Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library. *Adv. Funct. Mater.* **2015**, *25*, 6927–6935. https://doi.org/10.1002/adfm.201502811.
8. Kenry. Machine Learning-Assisted Clustering of Nanoparticle-Binding Peptides and Prediction of Their Properties. *Adv. Theory Simul.* **2023**, *6*, 2300122. https://doi.org/10.1002/adts.202300122.
9. Dias, L.M.S.; Fu, L.; Pereira, R.F.P.; et al. Evolving Photonic Authentication with Sustainable Luminescent Smart E-tags. *FlexMat* **2024**, *1*, 116–126. https://doi.org/10.1002/flm2.16.
10. Baig, M.M.; Khan, S.A.; Ahmad, H.; et al. 3D Printing of Hydrogels for Flexible Micro-supercapacitors. *FlexMat* **2024**, *1*, 79–99. https://doi.org/10.1002/flm2.14.
11. Ahmad, F.; Mahmood, A.; Muhmood, T. Machine Learning-Integrated Omics for the Risk and Safety Assessment of Nanomaterials. *Biomater. Sci.* **2021**, *9*, 1598–1608. https://doi.org/10.1039/D0BM01672A.
12. Ji, Z.; Guo, W.; Wood, E.L.; et al. Machine Learning Models for Predicting Cytotoxicity of Nanomaterials. *Chem. Res. Toxicol.* **2022**, *35*, 125–139. https://doi.org/10.1021/acs.chemrestox.1c00310.
13. Suwardi, A.; Wang, F.; Xue, K.; et al. Machine Learning-Driven Biomaterials Evolution. *Adv. Mater.* **2022**, *34*, 2102703. https://doi.org/10.1002/adma.202102703.

14. Wang, L.; Wang, H.; Bai, M.; et al. A Comparative Review: Research in Safety and Sustainability of Carbon Nanomaterials Without and With Machine Learning Assistance. *IEEE Access* **2024**, *12*, 167120–167152. https://doi.org/10.1109/ACCESS.2024.3494549.
15. Dang, Y.; Wang, G.; Su, G.; et al. Rational Construction of a Ni/$CoMoO_4$ Heterostructure with Strong Ni–O–Co Bonds for Improving Multifunctional Nanozyme Activity. *ACS Nano* **2022**, *16*, 4536–4550. https://doi.org/10.1021/acsnano.1c11012.
16. Fernandez, M.; Bilić, A.; Barnard, A.S. Machine Learning and Genetic Algorithm Prediction of Energy Differences between Electronic Calculations of Graphene Nanoflakes. *Nanotechnology* **2017**, *28*, 38LT03. https://doi.org/10.1088/1361-6528/aa82e5.
17. Gupta, N.; Jayaraman, A. Computational Approach for Structure Generation of Anisotropic Particles (CASGAP) with Targeted Distributions of Particle Design and Orientational Order. *Nanoscale* **2023**, *15*, 14958–14970. https://doi.org/10.1039/D3NR02425C.
18. Kløve, M.; Sommer, S.; Iversen, B.B.; et al. A Machine-Learning-Based Approach for Solving Atomic Structures of Nanomaterials Combining Pair Distribution Functions with Density Functional Theory. *Adv. Mater.* **2023**, *35*, 2208220. https://doi.org/10.1002/adma.202208220.
19. Konstantopoulos, G.; Koumoulos, E.P.; Charitidis, C.A. Digital Innovation Enabled Nanomaterial Manufacturing; Machine Learning Strategies and Green Perspectives. *Nanomaterials* **2022**, *12*, 2646. https://doi.org/10.3390/nano12152646.
20. Rangel DaCosta, L.; Sytwu, K.; Groschner, C.K.; et al. A Robust Synthetic Data Generation Framework for Machine Learning in High-Resolution Transmission Electron Microscopy (HRTEM). *NPJ Comput. Mater.* **2024**, *10*, 1–11. https://doi.org/10.1038/s41524-024-01336-0.
21. Wan, K.; He, J.; Shi, X. Construction of High Accuracy Machine Learning Interatomic Potential for Surface/Interface of Nanomaterials—A Review. *Adv. Mater.* **2024**, *36*, 2305758. https://doi.org/10.1002/adma.202305758.
22. Guerrero-Rivera, R.; Godínez-Garcia, F.J.; Hayashi, T.; et al. Machine-Learning Driven STM Images Prediction of Doped/Defective Graphene: towards Optimized Tools for 2D Nanomaterials Characterization. *Comput. Mater. Sci.* **2024**, *242*, 113076. https://doi.org/10.1016/j.commatsci.2024.113076.
23. Kuznetsova, V.; Coogan, Á.; Botov, D.; et al. Expanding the Horizons of Machine Learning in Nanomaterials to Chiral Nanostructures. *Adv. Mater.* **2024**, *36*, 2308912. https://doi.org/10.1002/adma.202308912.
24. Yang, L.; Wang, H.; Leng, D.; et al. Machine Learning Applications in Nanomaterials: Recent Advances and Future Perspectives. *Chem. Eng. J.* **2024**, *500*, 156687. https://doi.org/10.1016/j.cej.2024.156687.
25. Zhang, H.; Yang, M.; Wu, Q.; et al. Engineering Two-Dimensional Nanomaterials for Photothermal Therapy. *Angew. Chem. Int. Ed.* **2025**, *64*, e202424768. https://doi.org/10.1002/anie.202424768.
26. Cao, L.; Li, C.; Mueller, T. The Use of Cluster Expansions to Predict the Structures and Properties of Surfaces and Nanostructured Materials. *J. Chem. Inf. Model.* **2018**, *58*, 2401–2413. https://doi.org/10.1021/acs.jcim.8b00413.
27. Diao, S.; Wu, Q.; Li, S.; et al. From Synthesis to Properties: Expanding the Horizons of Machine Learning in Nanomaterials Research. *Mater. Horiz.* **2025**, *12*, 4133–4164. https://doi.org/10.1039/D4MH01909A.
28. Wang, M.; Wang, T.; Cai, P.; et al. Nanomaterials Discovery and Design through Machine Learning. *Small Methods* **2019**, *3*, 1900025. https://doi.org/10.1002/smtd.201900025.
29. Chen, R.; Liu, F.; Tang, Y.; et al. Combined First-Principles and Machine Learning Study of the Initial Growth of Carbon Nanomaterials on Metal Surfaces. *Appl. Surf. Sci.* **2022**, *586*, 152762. https://doi.org/10.1016/j.apsusc.2022.152762.
30. Alkharisi, M.K.; Dahish, H.A.; et al. Prediction Models for the Hybrid Effect of Nano Materials on Radiation Shielding Properties of Concrete Exposed to Elevated Temperatures. *Case Stud. Constr. Mater.* **2024**, *21*, e03750. https://doi.org/10.1016/j.cscm.2024.e03750.
31. Dhoble, S.; Wu, T.-H. Kenry Decoding Nanomaterial-Biosystem Interactions through Machine Learning. *Angew. Chem. Int. Ed.* **2024**, *63*, e202318380. https://doi.org/10.1002/anie.202318380.
32. Gao, X.J.; Yan, J.; Zheng, J.-J.; et al. Clear-Box Machine Learning for Virtual Screening of 2D Nanozymes to Target Tumor Hydrogen Peroxide. *Adv. Healthc. Mater.* **2023**, *12*, 2202925. https://doi.org/10.1002/adhm.202202925.
33. Zong, X.; Xu, X.; Pang, D.-W.; et al. Fine-Tuning Electron Transfer for Nanozyme Design. *Adv. Healthc. Mater.* **2025**, *14*, 2401836. https://doi.org/10.1002/adhm.202401836.
34. Ferdosi, S.; Stukalov, A.; Hasan, M.; et al. Enhanced Competition at the Nano–Bio Interface Enables Comprehensive Characterization of Protein Corona Dynamics and Deep Coverage of Proteomes. *Adv. Mater.* **2022**, *34*, 2206008. https://doi.org/10.1002/adma.202206008.
35. Winkler, D.A. Role of Artificial Intelligence and Machine Learning in Nanosafety. *Small* **2020**, *16*, 2001883. https://doi.org/10.1002/smll.202001883.
36. Firestein, K.L.; von Treifeldt, J.E.; Kvashnin, D.G.; et al. Young's Modulus and Tensile Strength of Ti_3C_2 MXene Nanosheets As Revealed by In Situ TEM Probing, AFM Nanomechanical Mapping, and Theoretical Calculations. *Nano Lett.* **2020**, *20*, 5900–5908. https://doi.org/10.1021/acs.nanolett.0c01861.

37. Lanjan, A.; Moradi, Z.; Srinivasan, S. Computational Framework Combining Quantum Mechanics, Molecular Dynamics, and Deep Neural Networks to Evaluate the Intrinsic Properties of Materials. *J. Phys. Chem. A* **2023**, *127*, 6603–6613. https://doi.org/10.1021/acs.jpca.3c02887.
38. Mortazavi, B.; Rajabpour, A.; Zhuang, X.; et al. Exploring Thermal Expansion of Carbon-Based Nanosheets by Machine-Learning Interatomic Potentials. *Carbon.* **2022**, *186*, 501–508. https://doi.org/10.1016/j.carbon.2021.10.059.
39. Li, J.; Telychko, M.; Yin, J.; et al. Machine Vision Automated Chiral Molecule Detection and Classification in Molecular Imaging. *J. Am. Chem. Soc.* **2021**, *143*, 10177–10188. https://doi.org/10.1021/jacs.1c03091.
40. Fjodorova, N.; Novič, M.; Venko, K.; et al. Cheminformatics and Machine Learning Approaches to Assess Aquatic Toxicity Profiles of Fullerene Derivatives. *IJMS* **2023**, *24*, 14160. https://doi.org/10.3390/ijms241814160.
41. Gao, W.; Yu, C.; Chen, R. Artificial Intelligence Accelerators Based on Graphene Optoelectronic Devices. *Adv. Photonics Res.* **2021**, *2*, 2100048. https://doi.org/10.1002/adpr.202100048.
42. Jin, W.; Pei, J.; Xie, P.; et al. Machine Learning-Based Prediction of Mechanical Properties and Performance of Nickel–Graphene Nanocomposites Using Molecular Dynamics Simulation Data. *ACS Appl. Nano Mater.* **2023**, *6*, 12190–12199. https://doi.org/10.1021/acsanm.3c01919.
43. Khot, A.C.; Dongale, T.D.; Nirmal, K.A.; et al. Amorphous Boron Nitride Memristive Device for High-Density Memory and Neuromorphic Computing Applications. *ACS Appl. Mater. Interfaces* **2022**, *14*, 10546–10557. https://doi.org/10.1021/acsami.1c23268.
44. Zhang, C.; Yang, B.; Peng, Z.; et al. Machine Learning-Based Prediction of Mechanical Properties of N-Doped γ-Graphdiyne. *Sci. China Mater.* **2024**, *67*, 1129–1139. https://doi.org/10.1007/s40843-023-2733-7.
45. Du, D.; Zhang, Y.; Li, X.; et al. First-Principles Calculations, Machine Learning and Monte Carlo Simulations of the Magnetic Coercivity of FexCo1−x Bulks and Nanoclusters. *Nanomaterials* **2025**, *15*, 577. https://doi.org/10.3390/nano15080577.
46. Kang, J.; Noh, S.H.; Hwang, J.; et al. First-Principles Database Driven Computational Neural Network Approach to the Discovery of Active Ternary Nanocatalysts for Oxygen Reduction Reaction. *Phys. Chem. Chem. Phys.* **2018**, *20*, 24539–24544. https://doi.org/10.1039/C8CP03801E.
47. Zhou, P.; Wang, M.; Tang, F.; et al. Machine Learning Accelerates the Screening of Efficient Metal-Oxide Catalysts for Photocatalytic Water Splitting. *Mater. Res. Bull.* **2024**, *179*, 112956. https://doi.org/10.1016/j.materresbull.2024.112956.
48. Pinho, B.; Torrente-Murciano, L. Dial-A-Particle: Precise Manufacturing of Plasmonic Nanoparticles Based on Early Growth Information—Redefining Automation for Slow Material Synthesis. *Adv. Energy Mater.* **2021**, *11*, 2100918. https://doi.org/10.1002/aenm.202100918.
49. Guo, G.; Goldfeder, J.; Lan, L.; et al. Towards End-to-End Structure Determination from x-Ray Diffraction Data Using Deep Learning. *NPJ Comput. Mater.* **2024**, *10*, 1–12. https://doi.org/10.1038/s41524-024-01401-8.
50. Guo, G.; Saidi, T.L.; Terban, M.W.; et al. Ab Initio Structure Solutions from Nanocrystalline Powder Diffraction Data via Diffusion Models. *Nat. Mater.* **2025**. https://doi.org/10.1038/s41563-025-02220-y.
51. Ma, K.; Gong, Y.; Aubert, T.; et al. Self-Assembly of Highly Symmetrical, Ultrasmall Inorganic Cages Directed by Surfactant Micelles. *Nature* **2018**, *558*, 577–580. https://doi.org/10.1038/s41586-018-0221-0.
52. Nishitsuji, R.; Nakashima, T.; Hisamoto, H.; et al. Simultaneous Recognition and Detection of Adenosine Phosphates by Machine Learning Analysis for Surface-Enhanced Raman Scattering Spectral Data. *Sensors* **2024**, *24*, 6648. https://doi.org/10.3390/s24206648.
53. Yu, Y.; Lu, W.; Yao, X.; et al. Machine Learning-Integrated Surface-Enhanced Raman Spectroscopy Analysis of Multicomponent Dye Mixtures. *Spectrochim. Acta Part. A Mol. Biomol. Spectrosc.* **2025**, *332*, 125806. https://doi.org/10.1016/j.saa.2025.125806.
54. Ieracitano, C.; Mammone, N.; Paviglianiti, A.; et al. A Conditional Generative Adversarial Network and Transfer Learning-Oriented Anomaly Classification System for Electrospun Nanofibers. *Int. J. Neur. Syst.* **2022**, *32*, 2250054. https://doi.org/10.1142/S012906572250054X.
55. Li, S.; Barnard, A.S. Safety-by-Design Using Forward and Inverse Multi-Target Machine Learning. *Chemosphere* **2022**, *303*, 135033. https://doi.org/10.1016/j.chemosphere.2022.135033.
56. Li, Y.; Liu, Y.; Miao, Y.; et al. Development of Heat-Resistant Tunnel Muck-Based Shotcrete for Geothermal Environments: Dual Drive of Combining Explainable Machine Learning and Microstructure Characterization. *Constr. Build. Mater.* **2025**, *473*, 140994. https://doi.org/10.1016/j.conbuildmat.2025.140994.
57. Tao, X. Compressive Strength Prediction of Nano-Modified Concrete: A Comparative Study of Advanced Machine Learning Techniques. *AIP Adv.* **2024**, *14*, 075017. https://doi.org/10.1063/5.0214890.
58. Ghorbani, K.; Mirchi, P.; Arabha, S.; et al. Lattice Thermal Conductivity and Young's Modulus of XN4 (X = Be, Mg and Pt) 2D Materials Using Machine Learning Interatomic Potentials. *Phys. Chem. Chem. Phys.* **2023**, *25*, 12923–12933. https://doi.org/10.1039/D3CP00746D.
59. Lindsey, R.K.; Goldman, N.; Fried, L.E.; et al. Chemistry-Mediated Ostwald Ripening in Carbon-Rich C/O Systems at Extreme Conditions. *Nat. Commun.* **2022**, *13*, 1424. https://doi.org/10.1038/s41467-022-29024-x.

60. Isfeldt, G.; Lundell, F.; Wohlert, J. Interaction of Complex Particles: A Framework for the Rapid and Accurate Approximation of Pair Potentials Using Neural Networks. *Phys. Rev. E* **2024**, *110*, 055305. https://doi.org/10.1103/PhysRevE.110.055305.
61. Pradeepa, A.; Arathi, P. Computing Degree-Based Topological Descriptors of Certain Tessellations of Kekulenes Using M-Polynomial and Neighborhood M-Polynomial. *Polycycl. Aromat. Compd.* **2025**, *45*, 36–59. https://doi.org/10.1080/10406638.2024.2384901.
62. Harper, D.R.; Nandy, A.; Arunachalam, N.; et al. Representations and Strategies for Transferable Machine Learning Improve Model Performance in Chemical Discovery. *J. Chem. Phys.* **2022**, *156*, 074101. https://doi.org/10.1063/5.0082964.
63. Dong, W.; Huang, Y.; Lehane, B.; et al. Multi-Objective Design Optimization for Graphite-Based Nanomaterials Reinforced Cementitious Composites: A Data-Driven Method with Machine Learning and NSGA-II. *Constr. Build. Mater.* **2022**, *331*, 127198. https://doi.org/10.1016/j.conbuildmat.2022.127198.
64. Gao, H.; Zhong, S.; Dangayach, R.; et al. Understanding and Designing a High-Performance Ultrafiltration Membrane Using Machine Learning. *Environ. Sci. Technol.* **2023**, *57*, 17831–17840. https://doi.org/10.1021/acs.est.2c05404.
65. Furxhi, I.; Roberts, S.; Cross, R.; et al. Bayesian Network Modelling for Predicting the Environmental Hazard of Silver Nanomaterials in Soils. *NanoImpact* **2025**, *37*, 100553. https://doi.org/10.1016/j.impact.2025.100553.
66. Kelkar, A.S.; Dallin, B.C.; Van Lehn, R.C. Identifying Nonadditive Contributions to the Hydrophobicity of Chemically Heterogeneous Surfaces via Dual-Loop Active Learning. *J. Chem. Phys.* **2022**, *156*, 024701. https://doi.org/10.1063/5.0072385.
67. Zhuang, Z.; Xu, Q.; Zeng, H.; et al. A Deep-Learning-Based Compact Method for Accelerating the Electrowetting Lattice Boltzmann Simulations. *Phys. Fluids* **2024**, *36*, 043323. https://doi.org/10.1063/5.0206608.
68. Wang, Z.; Ranasinghe, J.C.; Wu, W.; et al. Machine Learning Interpretation of Optical Spectroscopy Using Peak-Sensitive Logistic Regression. *ACS Nano* **2025**, *19*, 15457–15473. https://doi.org/10.1021/acsnano.4c16037.
69. Yu, H.; Zhou, G.-Y.; Liu, Y.-B.; et al. Deep Learning-Assisted Superhydrophobic LIG/MWCNT Wearable Sensor for Underwater Motion Detection. *IEEE Sens. J.* **2024**, *24*, 29392–29399. https://doi.org/10.1109/JSEN.2024.3434948.
70. Bruefach, A.; Ophus, C.; Scott, M.C. Analysis of Interpretable Data Representations for 4D-STEM Using Unsupervised Learning. *Microsc. Microanal.* **2022**, *28*, 1998–2008. https://doi.org/10.1017/S1431927622012259.
71. Ieracitano, C.; Paviglianiti, A.; Campolo, M.; et al. A Novel Automatic Classification System Based on Hybrid Unsupervised and Supervised Machine Learning for Electrospun Nanofibers. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 64–76. https://doi.org/10.1109/JAS.2020.1003387.
72. Sun, Z.; Shi, J.; Wang, J.; et al. A Deep Learning-Based Framework for Automatic Analysis of the Nanoparticle Morphology in SEM/TEM Images. *Nanoscale* **2022**, *14*, 10761–10772. https://doi.org/10.1039/D2NR01029A.
73. Kho, Z.; Bridger, A.; Butler, K.; et al. On the Use of Clustering Workflows for Automated Microstructure Segmentation of Analytical STEM Datasets. *APL Mater.* **2025**, *13*, 010901. https://doi.org/10.1063/5.0246329.
74. Li, Y.; Wang, Y.; Qi, S.; et al. Predicting Scattering from Complex Nano-Structures via Deep Learning. *IEEE Access* **2020**, *8*, 139983–139993. https://doi.org/10.1109/ACCESS.2020.3012132.
75. Boiko, D.A.; Kashin, A.S.; Sorokin, V.R.; et al. Analyzing Ionic Liquid Systems Using Real-Time Electron Microscopy and a Computational Framework Combining Deep Learning and Classic Computer Vision Techniques. *J. Mol. Liq.* **2023**, *376*, 121407. https://doi.org/10.1016/j.molliq.2023.121407.
76. Gu, Z.; Zhu, R.; Shen, T.; et al. Autonomous Nanorobots with Powerful Thrust under Dry Solid-Contact Conditions by Photothermal Shock. *Nat. Commun.* **2023**, *14*, 7663. https://doi.org/10.1038/s41467-023-43433-6.
77. Gandhi, A.M.; Shanmugan, S.; Gorjian, S.; et al. Performance Enhancement of Stepped Basin Solar Still Based on OSELM with Traversal Tree for Higher Energy Adaptive Control. *Desalination* **2021**, *502*, 114926. https://doi.org/10.1016/j.desal.2020.114926.
78. Jiang, T.; Gai, S.; Yin, Y.; et al. A Light/Thermal Cascaded-Driven Equipment for Machine Recognition Inspired by Water Lilies Using as Multifunctional Soft Actuator. *Chem. Eng. J.* **2024**, *495*, 153348. https://doi.org/10.1016/j.cej.2024.153348.
79. Guo, W.; Ma, Z.; Chen, Z.; et al. Thin and Soft Ti_3C_2Tx MXene Sponge Structure for Highly Sensitive Pressure Sensor Assisted by Deep Learning. *Chem. Eng. J.* **2024**, *485*, 149659. https://doi.org/10.1016/j.cej.2024.149659.
80. Zhang, J.; Perrin, M.L.; Barba, L.; et al. High-Speed Identification of Suspended Carbon Nanotubes Using Raman Spectroscopy and Deep Learning. *Microsyst. Nanoeng.* **2022**, *8*, 1–9. https://doi.org/10.1038/s41378-022-00350-w.
81. Zhao, Z.; Yang, X.; Wang, Y.; et al. Putting the Incoming/Outgoing Correlation (INOUTCO) Ion Imaging Surface Scattering Technique to the Test in O Atom Scattering from Graphite. *J. Phys. Chem. C* **2025**, *129*, 722–731. https://doi.org/10.1021/acs.jpcc.4c06578.
82. Hai, T.; Dahan, F.; Dhahad, H.A.; et al. Deep-Learning Optimization and Environmental Assessment of Nanomaterial's Boosted Hydrogen and Power Generation System Combined with SOFC. *Int. J. Hydrog. Energy* **2024**, *52*, 202–215. https://doi.org/10.1016/j.ijhydene.2022.11.332.
83. Jia, P.; Cao, C.; Lu, X.; et al. Machine Learning-Integrated Numerical Simulation for Predicting Photothermal Conversion Performance of Metallic Nanofluids. *Small* **2025**, *21*, 2408984. https://doi.org/10.1002/smll.202408984.

84. Balraadjsing, S.; Peijnenburg, W.J.G.M.; Vijver, M.G. Exploring the Potential of in Silico Machine Learning Tools for the Prediction of Acute Daphnia Magna Nanotoxicity. *Chemosphere* **2022**, *307*, 135930. https://doi.org/10.1016/j.chemosphere.2022.135930.
85. Guo, H.; Lesani, P.; Zreiqat, H.; et al. A Fluorescent Sensor Array Based on Carbon Dots for the Accurate Determination of pH. *Sens. Diagn.* **2024**, *3*, 1923–1934. https://doi.org/10.1039/D4SD00275J.
86. Okeke, C.; Juma, I.; Cobarrubia, A.; et al. Probing Anharmonic Phonons in WS2 van Der Waals Crystal by Raman Spectroscopy and Machine Learning. *iScience* **2023**, *26*, 107174. https://doi.org/10.1016/j.isci.2023.107174.
87. Exner, T.E.; Papadiamantis, A.G.; Melagraki, G.; et al. Metadata Stewardship in Nanosafety Research: Learning from the Past, Preparing for an "on-the-Fly" FAIR Future. *Front. Phys.* **2023**, *11*, 1233879. https://doi.org/10.3389/fphy.2023.1233879.
88. He, S.; Nader, K.; Abarrategi, J.S.; et al. NANO.PTML Model for Read-across Prediction of Nanosystems in Neurosciences. Computational Model and Experimental Case of Study. *J. Nanobiotechnol.* **2024**, *22*, 435. https://doi.org/10.1186/s12951-024-02660-9.
89. Cruz, C.; Matatagui, D.; Ramírez, C.; et al. Carbon SH-SAW-Based Electronic Nose to Discriminate and Classify Sub-Ppm NO_2. *Sensors* **2022**, *22*, 1261. https://doi.org/10.3390/s22031261.
90. Shao, S.; Xie, C.; Xia, Y.; et al. Highly Conjugated Three-Dimensional van Der Waals Heterostructure-Based Nanocomposite Films for Ultrahigh-Responsive TEA Gas Sensors at Room Temperature. *J. Mater. Chem. A* **2022**, *10*, 2995–3008. https://doi.org/10.1039/D1TA09749K.
91. Singh, S.; Saggu, I.S.; Singh, S.; et al. Detection of DMF and NH_3 at Room Temperature Using a Sensor Based on a MoS_2 /Single-Walled Carbon Nanotube Composite. *ACS Appl. Nano Mater.* **2023**, *6*, 10698–10712. https://doi.org/10.1021/acsanm.3c01638.
92. Jeindl, A.; Domke, J.; Hörmann, L.; et al. Nonintuitive Surface Self-Assembly of Functionalized Molecules on Ag(111). *ACS Nano* **2021**, *15*, 6723–6734. https://doi.org/10.1021/acsnano.0c10065.
93. Leppänen, E.; Aarva, A.; Sainio, S.; et al. Connection between the Physicochemical Characteristics of Amorphous Carbon Thin Films and Their Electrochemical Properties. *J. Phys. Condens. Matter* **2021**, *33*, 434002. https://doi.org/10.1088/1361-648X/ac1a2e.
94. Packwood, D.M. Bi-Functional On-Surface Molecular Assemblies Predicted from a Multifaceted Computational Approach. *Adv. Phys. Res.* **2022**, *1*, 2200019. https://doi.org/10.1002/apxr.202200019.
95. Khan, S.A.; Farooq, U.; Imran, M.; et al. Mathematical and Artificial Neural Network Modeling to Predict the Heat Transfer of Mixed Convective Electroosmotic Nanofluid Flow with Helmholtz-Smoluchowski Velocity and Multiple Slip Effects: An Application of Soft Computing. *Case Stud. Therm. Eng.* **2024**, *61*, 104950. https://doi.org/10.1016/j.csite.2024.104950.
96. Zhang, Y.; Li, Q.; Feng, W.; et al. Regulating Electron Transfer in Vanadium-Based Metal–Organic Frameworks via the Synergy of Linker Engineering and Machine Learning for Efficient and Reversible Aqueous Zinc Ion Batteries. *Adv. Mater.* **2025**, *37*, 2507609. https://doi.org/10.1002/adma.202507609.
97. Li, Q.; Zhang, Y.; Feng, W.; et al. Manganese–Based Metal–Organic Coordination for Aqueous Zinc–Ion Batteries with Varying Mechanical Adaptability and Machine Learning–Assisted Performance Decoding. *Adv. Mater.* **2025**, *37*, 2507951. https://doi.org/10.1002/adma.202507951.
98. Aytaç, E.; Khanzada, N.K.; Ibrahim, Y.; et al. Reverse Osmosis Membrane Engineering: Multidirectional Analysis Using Bibliometric, Machine Learning, Data, and Text Mining Approaches. *Membranes* **2024**, *14*, 259. https://doi.org/10.3390/membranes14120259.
99. Baig, N.; Usman, J.; Abba, S.I.; et al. Fractionation of Dyes/Salts Using Loose Nanofiltration Membranes: Insight from Machine Learning Prediction. *J. Clean. Prod.* **2023**, *418*, 138193. https://doi.org/10.1016/j.jclepro.2023.138193.
100. Li, J.; Meng, K.; Yu, X.; et al. Mechanistic Insight into a Graphene-like Stimulus-Responsive Desalination Membrane from Molecular Dynamics and First Principles. *Diam. Relat. Mater.* **2023**, *136*, 109910. https://doi.org/10.1016/j.diamond.2023.109910.
101. Madejski, G.R.; Ahmad, S.D.; Musgrave, J.; et al. Silicon Nanomembrane Filtration and Imaging for the Evaluation of Microplastic Entrainment along a Municipal Water Delivery Route. *Sustainability* **2020**, *12*, 10655. https://doi.org/10.3390/su122410655.
102. Zhang, M.; He, H.; Huang, Y.; et al. Machine Learning Integrated High Quantum Yield Blue Light Carbon Dots for Real-Time and on-Site Detection of Cr(VI) in Groundwater and Drinking Water. *Sci. Total Environ.* **2023**, *904*, 166822. https://doi.org/10.1016/j.scitotenv.2023.166822.
103. Ji, Y.; Ma, S.; Lv, S.; et al. Nanomaterials for Targeted Delivery of Agrochemicals by an All-in-One Combination Strategy and Deep Learning. *ACS Appl. Mater. Interfaces* **2021**, *13*, 43374–43386. https://doi.org/10.1021/acsami.1c11914.
104. Hao, T.; Zhou, H.; Gai, P.; et al. Deep Learning-Assisted Single-Atom Detection of Copper Ions by Combining Click Chemistry and Fast Scan Voltammetry. *Nat. Commun.* **2024**, *15*, 10292. https://doi.org/10.1038/s41467-024-54743-8.
105. Stuart, S.; Watchorn, J.; Gu, F.X. An Interpretable Machine Learning Framework for Modelling Macromolecular Interaction Mechanisms with Nuclear Magnetic Resonance. *Digit. Discov.* **2023**, *2*, 1697–1709. https://doi.org/10.1039/D3DD00009E.

106. Sun, H.; Amin, M.N.; Qadir, M.T.; et al. Investigating the Effectiveness of Carbon Nanotubes for the Compressive Strength of Concrete Using AI-Aided Tools. *Case Stud. Constr. Mater.* **2024**, *20*, e03083. https://doi.org/10.1016/j.cscm.2024.e03083.
107. Della Pia, F.; Zen, A.; Kapil, V.; et al. On the Increase of the Melting Temperature of Water Confined in One-Dimensional Nano-Cavities. *J. Chem. Phys.* **2024**, *161*, 224706. https://doi.org/10.1063/5.0239452.
108. Vakharia, V.; Castelli, I.E.; Bhavsar, K.; et al. Bandgap Prediction of Metal Halide Perovskites Using Regression Machine Learning Models. *Phys. Lett. A* **2022**, *422*, 127800. https://doi.org/10.1016/j.physleta.2021.127800.
109. Chen, K.; Li, N.; Luo, Y.; et al. High-Performance Hardware Primitives Based on Sub-10 Nm Nanodiodes for Cryptography Applications. *J. Mater. Chem. C* **2024**, *12*, 17878–17889. https://doi.org/10.1039/D4TC02206H.
110. Wang, S.; Zhu, J.; Blackwell, R.; et al. Automated Tip Conditioning for Scanning Tunneling Spectroscopy. *J. Phys. Chem. A* **2021**, *125*, 1384–1390. https://doi.org/10.1021/acs.jpca.0c10731.
111. Guccione, P.; Diacono, D.; Toso, S.; et al. Towards the Extraction of the Crystal Cell Parameters from Pair Distribution Function Profiles. *IUCrJ* **2023**, *10*, 610–623. https://doi.org/10.1107/S2052252523006887.
112. He, J.; Wang, C.; Tang, H.; et al. Prospective Research on the Tribological Behavior of Graphdiyne Nanofluid and Its Machine Learning Performance Prediction. *Appl. Surf. Sci.* **2025**, *696*, 162954. https://doi.org/10.1016/j.apsusc.2025.162954.