PCCP

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2026, **28**, 5072

Feature engineering methods for machine learning in heterogeneous catalysis

Yu Jin,[a] Hang-Biao Lv,[a] Shisheng Zheng*[a] and Jian-Feng Li *[ab]

Received 11th November 2025,
Accepted 23rd January 2026

DOI: 10.1039/d5cp04352b

rsc.li/pccp

Machine learning is revolutionizing the field of heterogeneous catalysis, transitioning from a supporting tool to a central force in materials discovery and mechanistic understanding. At the heart of this transformation lies feature engineering, which bridges the catalyst structure with predictive modeling capabilities. In this review, we provide a systematic overview of the evolution of feature engineering in heterogeneous catalysis. This progression spans hand-crafted descriptors, symbolic regression methods, graph-based features that capture intricate chemical and geometric relationships, topological data features encoding multiscale structural invariants, and most recently, multimodal representations that integrate textual data and structure into unified feature spaces. Despite these advancements, several challenges remain in feature engineering, including the underdevelopment of multimodal representations, limited model interpretability, and the absence of cross-scale structural descriptors. Emerging strategies aimed at addressing these issues are discussed in detail. We hope that this review will inspire further innovation in feature engineering methodologies tailored to the continued advancement of heterogeneous catalysis.

1. Introduction

Heterogeneous catalysis, a cornerstone of clean energy conversion and storage, underpins key processes such as water splitting for hydrogen production, CO_2 electroreduction (CO_2RR), and fuel cell reactions.[1–6] However, the intrinsic complexity of catalytic systems, which arises from high-dimensional chemical spaces, intricate interfacial reaction pathways, and multiscale physicochemical coupling, makes traditional trial-and-error experiments and standalone simulations inefficient for rational catalyst design and mechanistic understanding. In this context, machine learning (ML) has emerged as a transformative approach, capable of autonomously learning complex structure–performance relationships from growing volumes of computational and experimental data.[7,8] It enables rapid prediction of adsorption energies,[9] efficient catalyst screening,[10,11] and discovery of nonlinear correlations among multiple variables.[12,13] ML has evolved from a mere auxiliary tool into a central paradigm for materials discovery and mechanistic elucidation in heterogeneous catalysis.[14] Consequently, artificial intelligence (AI) is increasingly viewed not as a supporting technique but as a driving force reshaping catalytic science.

In ML applications for heterogeneous catalysis, model performance, whether predicting catalytic activity, adsorption energies, or material stability, critically depends on the features used to represent the system.[15–17] Features are numerical encodings of key physicochemical properties that bridge the catalyst structure with predictive models.[18–22] This involves selecting, transforming, and constructing descriptors that capture essential determinants of catalytic behavior, such as electronic structure, atomic geometry, and composition. Well-designed features enhance predictive accuracy, generalization, and computational efficiency by focusing on the most informative variables.[23] For example, Vinchurkar *et al.* found that "effective coordination number" and the catalyst's "catalyst electronegativity" were the most important features in their model, and through symbolic regression deduced that the adsorption energy is approximately proportional to the square of the catalyst electronegativity.[24] Consequently, feature engineering is not merely a preprocessing step but a decisive factor shaping the reliability and interpretability of ML-driven catalysis research.

This review traces the evolution of feature engineering as a central driver of ML in heterogeneous catalysis with a primary focus on how to encode catalyst structures. Early studies relied on handcrafted features such as electronegativity, atomic radius, and other fundamental properties, forming the basis for high-throughput screening using density functional theory (DFT). The subsequent introduction of symbolic regression methods, exemplified by the sure independence screening

[a] *College of Energy, State Key Laboratory of Physical Chemistry of Solid Surfaces, iChEM, College of Chemistry and Chemical Engineering, Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, Fujian, China. E-mail: zhengss@xmu.edu.cn, Li@xmu.edu.cn*
[b] *Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen 361000, Fujian, China*

and sparsifying operator (SISSO), enabled automated discovery of low-dimensional, interpretable descriptors from vast candidate spaces. With the emergence of graph neural networks (GNNs), structural representation advanced toward end-to-end learning of atomic configurations and local chemical environments. More recently, topological data analysis (TDA) has provided mathematical tools such as persistent homology to quantify multiscale geometric and topological invariants, while multimodal fusion strategies have begun integrating structural, compositional, and textual information into unified representations. This review discusses the application of these feature paradigms to catalytic activity prediction and stability analysis. Finally, we outline emerging opportunities in developing more advanced multimodal feature representations, enhancing model interpretability, and establishing cross-scale feature engineering frameworks, all aimed at accelerating the rational design of high-performance catalytic systems.

2. Machine learning in catalysis research

AI is rapidly evolving from a supportive tool to a core driver in heterogeneous catalysis research. Its key value lies in extracting hidden "structure–performance relationships" from high-dimensional, complex, multi-source data. This enables accurate prediction of catalytic properties and rapid screening of high-performance materials. Such progress relies on a systematic ML workflow (Fig. 1), which continuously improves model performance and scientific insight through a data-driven closed loop.[16,25] The workflow encompasses several key stages. It begins with data collection and preparation, which integrates multi-source data from experimental measurements, literature databases, and high-throughput computations. Subsequent steps, including data cleaning, standardization, encoding, and augmentation, ensure data reliability and consistency. The processed data is then divided into training, validation, and test sets to lay the foundation for assessing model generalizability. A core component is the feature engineering, which aims to extract informative and physically/chemically interpretable feature representations from raw structural and property data. This stage typically involves three sub-steps: (1) feature construction: generating new descriptors based on domain knowledge, such as electronegativity, atomic radius, and coordination number-related descriptors; (2) feature selection: employing filter, wrapper, or embedded methods to identify the most predictive variables from the high-dimensional feature space; (3) feature transformation: optimizing the feature space structure through techniques like principal component analysis (PCA), nonlinear mapping, or discretization to enhance model stability and generalization. In this review, these methodological sub-steps are manifested through five primary feature paradigms: hand-crafted features, SISSO, graph based features, topological data features, and multimodal features. The process then proceeds to model training and evaluation. Algorithms such as neural networks, gradient boosting trees, and GNNs are applied to learn the feature-performance mapping. Model performance is assessed using multiple metrics (*e.g.*, RMSE, R^2, and MAE) to guide model selection. Following optimization, the model enters the deployment and application phase, where it is validated and

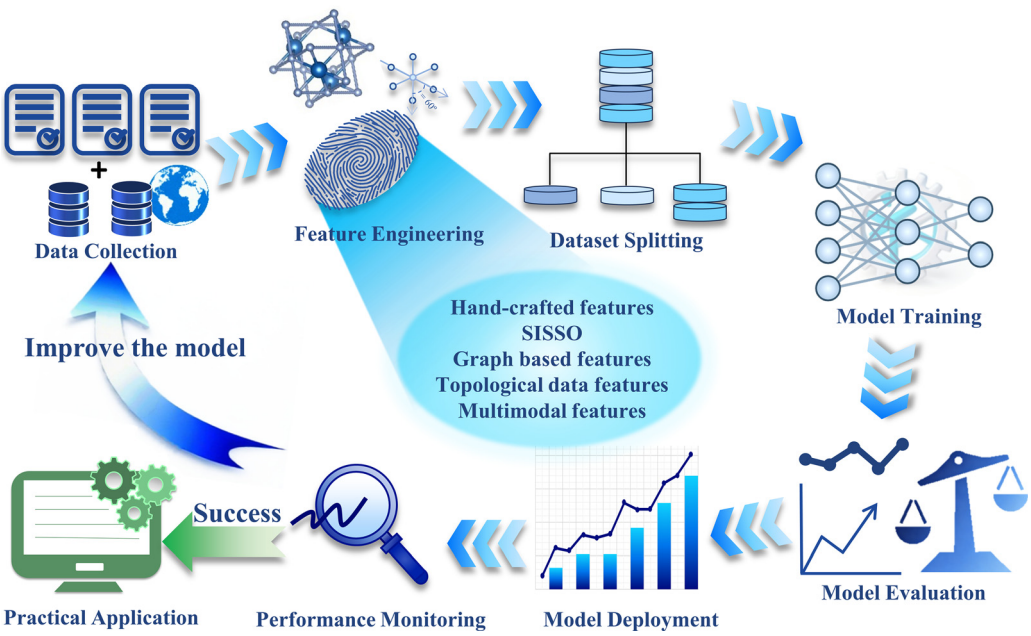

Fig. 1 Workflow for the development and deployment of machine learning models in heterogeneous catalysis. The process begins with data collection and proceeds through stages of feature engineering, dataset splitting, model training, evaluation, and deployment. Following deployment, continuous performance monitoring guides outcomes toward two critical pathways: successful models advance to practical application, while models requiring refinement initiate a cycle to improve the model, which informs subsequent rounds of data collection for iterative enhancement.

monitored in real-world catalytic systems. Finally, through continuous iteration and model refinement, a closed-loop optimization is achieved. This loop connects model predictions with experimental validation, mechanism elucidation, and knowledge feedback, thereby providing a robust foundation for the intelligent design of heterogeneous catalytic materials.

A key application of ML in electrocatalysis is the rapid prediction of performance metrics, including adsorption energies, activity, and selectivity, enabling high-throughput screening of candidate materials.[26–30] For example, Rosen *et al.* showed that the crystal graph convolutional neural network (CGCNN) can directly predict fundamental properties, such as band gaps, from crystal structures.[31] By incorporating descriptors such as intermediate adsorption energies, active-site coordination, and reaction conditions (*e.g.*, potential and pH), AI models can accurately predict product selectivity in complex reaction networks.[32] For the oxygen evolution reaction (OER), symbolic regression approaches like SISSO have been used to extract concise physical descriptors from large feature spaces, facilitating precise activity trend predictions.[33] In studies of g-CN-supported double-atom catalysts (DACs), Bian *et al.* employed feed-forward neural networks trained on DFT data to predict limiting potentials for various bimetallic combinations, efficiently identifying superior catalysts for the catalytic CO_2RR to CO and HCOOH.[28] More machine-learning-driven catalyst discovery studies can be found in previous review articles.[16,34–36]

Another key application of ML lies in machine-learning interatomic potentials (MLIPs), which enable atomistic simulations of catalysts with near-DFT accuracy at unprecedented computational efficiency.[37,38] MLIPs achieve this by learning high-dimensional potential energy surfaces directly from reference DFT data using flexible, data-driven representations such as neural networks, graph-based models, and other regression frameworks, thereby overcoming the functional limitations of classical empirical force fields and extending simulations to larger system sizes and longer timescales. To ensure the reliability and transferability of such potentials across vast configurational spaces, uncertainty-aware active learning frameworks have been developed as a systematic strategy to iteratively identify poorly sampled regions and enrich training datasets.[39,40] In the electrocatalysis domain, a growing body of research has shown that MLP-driven molecular dynamics can significantly reduce computational cost compared with conventional DFT simulations.[16,41,42] For instance, Lian *et al.* used high-accuracy machine-learning potential-driven molecular dynamics simulations to investigate oxide-derived copper electrocatalysts and showed that subsurface oxygen diffusion occurs over spatiotemporal scales extending from seconds to hours, a regime that is experimentally relevant but effectively unreachable by conventional DFT due to prohibitive computational cost.[43]

ML has emerged as an essential tool in heterogeneous catalysis, offering powerful capabilities for predicting material properties, adsorption energies, catalytic activity, and stability. It is fundamentally reshaping the paradigms of catalyst design and discovery, driving the field toward a more rational and data-driven future. Central to these advances is the construction of effective system representations and feature sets, with feature engineering serving as a critical link between underlying physicochemical mechanisms and ML models. The continued development of feature engineering is thus pivotal to the effectiveness and depth of AI applications in heterogeneous catalysis.

3. Feature engineering in heterogeneous catalysis

3.1. Rule-based or knowledge-based features

3.1.1. Hand-crafted features. Hand-crafted features are parameters designed by researchers based on physicochemical intuition and theoretical models to quantitatively characterize the intrinsic properties of catalysts. These features derive from elemental and compound properties, quantum chemical calculations, and experimental measurements, spanning multiple dimensions such as electronic structure, geometric configuration, thermodynamic stability, and surface chemistry.[44] For example, properties like electronegativity, atomic radius, valence electron count, and density can be directly obtained from databases, whereas quantities such as metallic d-band centers, adsorption free energies, and coordination numbers require computational evaluation or in-depth analysis. Hand-crafted features offer clear physical interpretation, linking catalytic performance to material properties and providing a theoretical basis for rational catalyst design. They encapsulate the "*a priori* knowledge" in materials science, bridging empirical exploration and predictive modeling in heterogeneous catalysis. This framework has evolved from single physicochemical descriptors to multi-parameter combinations, reflecting the progressive enrichment of feature representations.

Among numerous handcrafted features, the d-band center theory represents a paradigmatic example of successfully establishing a quantitative correlation between the electronic structure and catalytic activity.[45,46] By describing the relative position of metallic d-band centers with respect to the Fermi level, this theory provides a mechanistic understanding of how adsorption strengths of reaction intermediates are regulated on transition metal surfaces. Originating from Newns *et al.*'s quantum model[47] and systematically elaborated by Nørskov *et al.*,[48] the d-band center theory has undergone continuous refinement. It has been employed to rationalize adsorption and reactivity variations across different metal surfaces,[49] guide the design of alloy catalysts,[50] and elucidate the influence of strain or surface modification on catalytic performance,[51] ultimately giving rise to the well-known "volcano plot" for predicting activity trends.[52] Beyond the d-band center itself, additional descriptors such as d-bandwidth, filling factor, and coupling matrix elements, introduced by Nilsson[53] and Ruban *et al.*,[50,54] have been incorporated to enhance theoretical precision. Experimentally, Stamenkovic *et al.* demonstrated that the formation of a Pt-skin structure on $Pt_3Ni(111)$ lowers the d-band center of surface Pt by approximately 0.34 eV, resulting in a tenfold

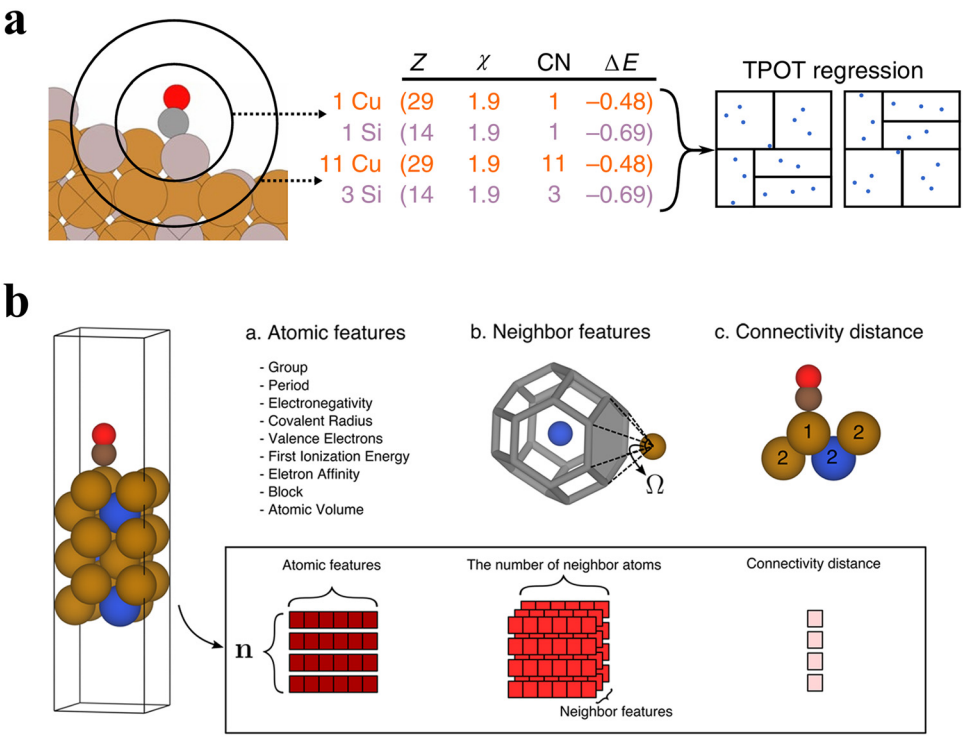

Fig. 2 (a) Fingerprint of the coordination site. Adsorption sites are reduced to numerical representations, or fingerprints, and these fingerprints are used as model features by TPOT[55] to predict ΔE_{CO}. Reprinted with permission.[44] Copyright 2018, Springer Nature. (b) Nine basic atomic properties are presented by one-hot encoding[56] to prepare the atomic feature vectors. Reprinted with permission.[57] Copyright 2019, American Chemical Society.

enhancement of the oxygen reduction reaction (ORR) activity relative to Pt(111). This work provides direct evidence of how electronic structure modulation can govern catalytic behavior, highlighting the predictive power and practical relevance of d-band-based descriptors in electrocatalysis.

As research in electrocatalysis has advanced, handcrafted features have gradually evolved from single-parameter descriptors to multi-parameter combinations forming systematic descriptor frameworks. A representative example is provided by Trand *et al.*, who integrated four categories of physicochemical descriptors: the atomic number of the element (Z), the Pauling electronegativity (χ), the coordination number of the element with the adsorbate (CN), and the median adsorption energy between the adsorbate and the pure element (ΔE), as illustrated in Fig. 2a. This approach constructs a 32-dimensional feature vector, offering a comprehensive digital representation of the local chemical environment at adsorption sites. By establishing effective mapping from simple elemental properties to complex catalytic performance, it enables accurate predictions of adsorption energies for the CO_2RR and the hydrogen evolution reaction (HER).[44] Building on this framework, hand-crafted features have been widely applied in machine-learning-assisted high-throughput screening of electrocatalysts.[58,59] For example, Back *et al.* further expanded the feature system (Fig. 2b) by incorporating additional atomic properties, including periodic table position, electronegativity, atomic volume, valence electron count, first ionization energy, electron affinity, and atomic radius. Moreover, by introducing a Voronoi polyhedron-based neighborhood solid angle descriptor, this enhanced feature set achieved an average absolute error of only 0.15 eV in predicting *CO and *H adsorption energies, thereby facilitating high-throughput screening across the vast catalyst design space.[57] In the study of single-atom and diatomic catalysts, handcrafted features have proven to be highly effective in elucidating structure–activity relationships. By carefully selecting descriptors such as d/p electron count, oxide formation enthalpy, and electronegativity, researchers successfully predicted the selectivity of single-atom catalysts (SACs) for H_2O_2 generation.[57] Similarly, active sites for the CO_2RR on dealloyed gold surfaces were identified,[60] and the critical role of interatomic spacing in governing HER activity within g-CN systems was established.[61] Subsequently, using hand-crafted geometric and electronic descriptors combined with random-forest models, a materials genome containing 279 bi-atom catalysts was constructed, from which 9 HER-active, 3 OER-active, and 5 ORR-active catalysts were high-throughput screened, with AuCo/g-CN identified as a rare trifunctional HER/OER/ORR catalyst.[62] Xu *et al.* using physically interpretable hand-crafted descriptors combined with an XGBoost model, screened 196 S/N-coordinated SACs and uncovered 17 promising NRR catalysts, among which Mo@S_3N_1 and W@S_3N_1 exhibited the best performance.[63]

To ensure robustness, many hand-crafted features are physics-informed and mathematically designed. Notably, the smooth overlap of atomic positions (SOAP) descriptor (Fig. 3) represents local atomic environments by expanding a smooth atomic neighbor density on the basis of radial functions and

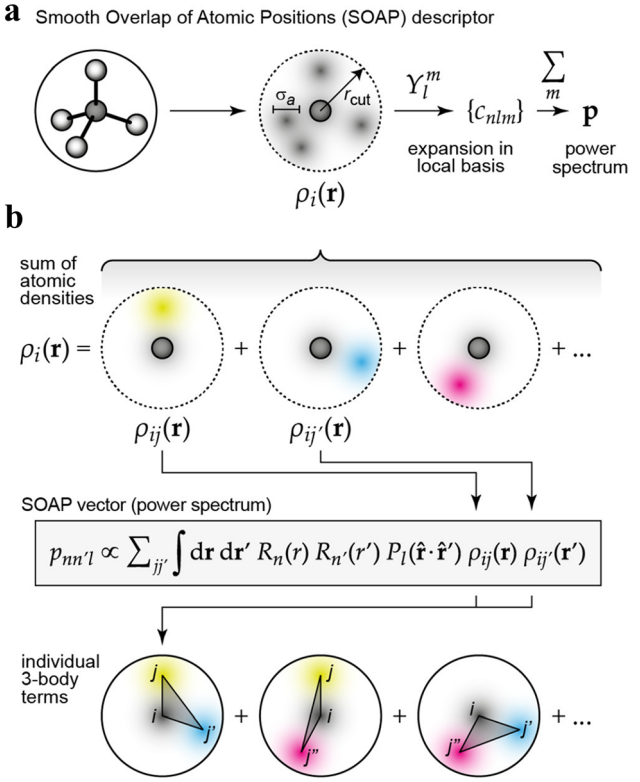

Fig. 3 (a) Illustration of the construction principle of the smooth overlap of atomic positions (SOAP) descriptor. First, the neighbor atomic density ρ around a central atom is expanded in a local basis composed of radial basis functions and spherical harmonics Y_l^m. Then, the expansion coefficients C_{nlm} are summed over the squared modulus in the m direction to obtain the power spectrum vector p, thereby ensuring the rotational invariance of the descriptor; (b) elucidation of this construction from a mathematical structural perspective. Reprinted with permission.[64] Copyright 2021, American Chemical Society.

spherical harmonics, yielding a continuous, high-dimensional representation that is invariant to translation, rotation, and permutation of identical atoms.[65] Among structural descriptors tested in kernel ridge regression models for hydrogen adsorption energies on MoS_2 and Cu–Au nanoclusters, Jäger *et al.* found that SOAP achieved the lowest mean absolute error in modeling HER activity.[66] In the screening of catalysts for CO_2 hydrogenation at complex metal–oxide interfaces, Nielsen *et al.* found that combining SOAP descriptors with the WWL-GPR model can efficiently predict catalyst adsorption energies.[67] When combined with sparse Gaussian process regression (SGPR), SOAP can be used to develop data-efficient machine-learned potentials with built-in uncertainty quantification, enabling active-learning-driven refinement of training datasets during molecular dynamics simulations.[68,69] To improve scalability over large configurational and chemical spaces, sparse Bayesian committee machine (BCM) schemes partition the descriptor space into multiple local SGPR experts and integrate their predictions within a Bayesian framework, preserving uncertainty estimates while reducing computational cost;[70] such BCM-based potentials have been explored as a route toward constructing transferable potentials spanning wide materials spaces, enabling high-throughput molecular dynamics simulations of multicomponent and multiphase systems.[70]

Manual feature design in heterogeneous catalysis faces intrinsic limitations. Expert-crafted descriptors often struggle to explore high-dimensional feature spaces, capture nonlinear multi-factor interactions, or generalize to novel materials such as high-entropy alloys, complex oxides, and metal–organic frameworks. These challenges have motivated interpretable automated feature engineering methods, such as SISSO, which efficiently identify optimal feature combinations from large pools of primary descriptors. By revealing subtle interactions inaccessible to human intuition, such approaches enhance descriptor discovery and enable rational design of complex catalytic systems.

3.1.2. SISSO. Interpretability is a fundamental requirement for ML-driven catalytic research.[71] The introduction of the SISSO method, originally proposed by Ouyang *et al.*, represents a significant advance in feature engineering, shifting from manual design to automated, interpretable descriptor discovery.[72] SISSO leverages symbolic regression to screen relevant descriptors from high-dimensional feature spaces and directly generates mathematically meaningful expressions, thereby establishing a systematic framework for constructing catalytic material descriptors. Interpretability is a fundamental requirement for ML-driven catalytic research.[71] The introduction of the SISSO method, originally proposed by Ouyang *et al.*, represents a significant advance in feature engineering, shifting from manual design to automated, interpretable descriptor discovery.[72] SISSO leverages symbolic regression to screen relevant descriptors from high-dimensional feature spaces and directly generates mathematically meaningful expressions, thereby establishing a systematic framework for constructing catalytic material descriptors. Before applying symbolic regression, domain-specific filtering is applied to prune the initial variable pool by enforcing dimensional consistency, removing redundant features through correlation analysis, and retaining only variables that are robustly available from DFT or experimental databases.

The fundamental framework of SISSO is illustrated in Fig. 4a. Its workflow consists of two sequential steps: first, ''sure independence screening'' rapidly reduces the candidate feature space; second, a ''sparsifying operator'' precisely identifies optimal low-dimensional descriptors.[72] By integrating compressive sensing with symbolic regression, SISSO enables the automated extraction of descriptors from complex expressions involving numerous primary features. The resulting descriptors are presented as analytical formulas, providing a transparent link between data-driven modeling and the underlying physical mechanisms.[72,73]

With methodological advancements, SISSO has demonstrated substantial applicability in the study of catalytic materials. For example, in predicting the relative stability of octahedral binary compounds, SISSO derived analytical formulas linking energy stability to complex feature spaces, thereby establishing clear mappings between material properties.[73]

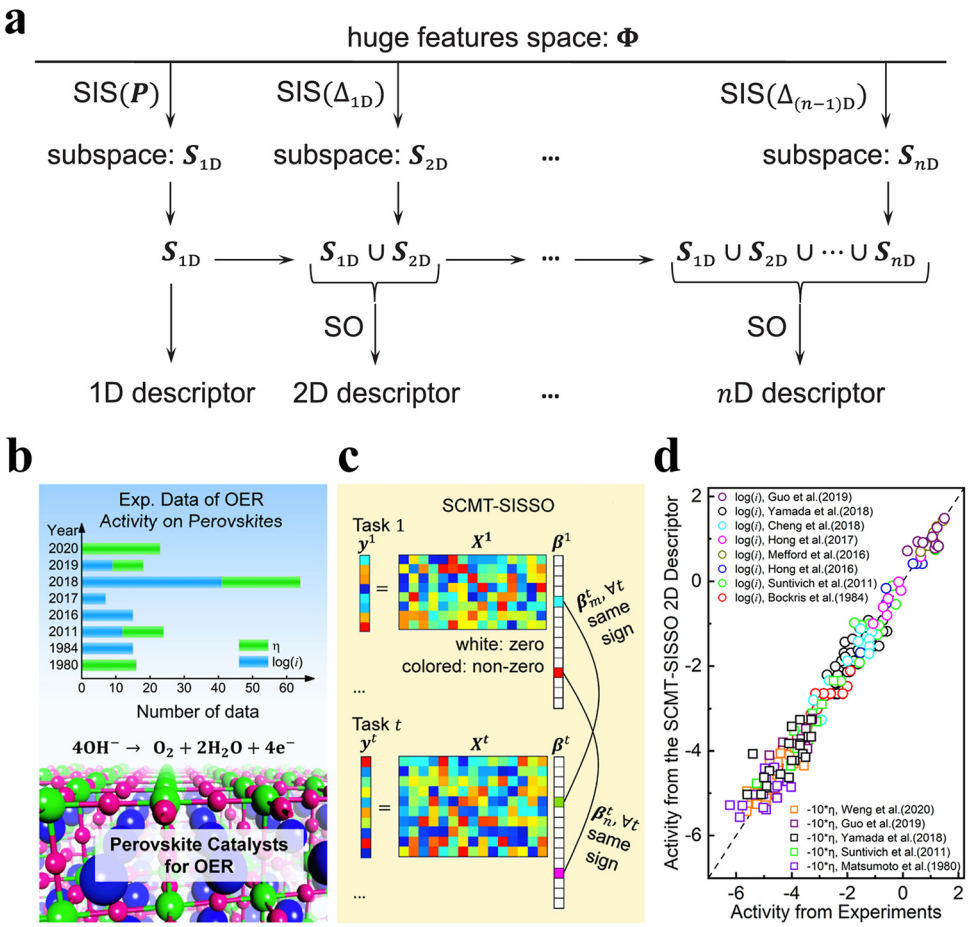

Fig. 4 (a) The method SISSO combines unified subspaces having the largest correlation with residual errors (or P) generated by sure independence screening (SIS) with a sparsifying operator (SO) to further extract the best descriptor. Reproduced with permission.[72] Copyright 2018, American Physical Society. (b) Distribution of the collected experimental OER activity data on perovskite catalysts in publications up to the year 2020 (the time of starting this project). (c) Idea of sign-constrained MTL. The *i*th coefficients βit in all *t* have the same sign. (d) Comparison of the activity data between the identified 2D descriptor (d_B, n_B) and the experiments. The colors denote the source of the data sets. Reproduced with permission.[74] Copyright 2023, American Chemical Society.

Building on this foundation, the method has been extensively applied to catalytic systems.[33] Wang *et al.* introduced the sign-constrained multi-task learning (SCMT-SISSO) framework (Fig. 4b and c), which addresses discrepancies in experimental data by enforcing sign consistency of descriptor coefficients across multiple sources. Using 182 data points from 13 independent studies, they identified an effective two-dimensional descriptor, (d_B, n_B), where d_B corresponds to the number of *d* electrons in the B-site metal and n_B denotes its oxidation state (Fig. 4d). This descriptor enabled the screening of 36 660 perovskite materials, successfully predicting several high-performance OER catalysts whose activity was subsequently validated experimentally.[74] In addition, Fung *et al.* employed compressed sensing to extract key descriptors for the HER reaction in SACs.[75] Similarly, in studies on selective alkene oxidation, Foppa *et al.* applied SISSO to reveal intrinsic correlations between key features and catalytic performance, based on 12 vanadium/manganese catalysts and 55 physicochemical parameters.[76]

To address challenges in practical applications, researchers have continued to optimize and extend the SISSO methodology to identify universal descriptors across diverse catalytic reactions. For example, Gong *et al.* proposed the physically meaningful feature engineering and selection (PFESS) framework, inspired by SISSO, and developed the ARSC descriptor with explicit physical interpretation, expressed analytically as $\Phi = (1 + k\alpha) \times \phi_{xy}$ [62]. As illustrated in Fig. 5, the descriptor construction follows a systematic four-step process: (i) establishing primary atomic properties (*A*) based on the d-band shape of homonuclear sites; (ii) selecting optimal parameters (*R*) by incorporating reactant effects; (iii) introducing heteronuclear intermetallic synergistic effects (*S*) *via* the PFESS framework; and (iv) integrating coordination environment influences (*C*) to form the final ARSC descriptor. This approach demonstrates how complex descriptors can be progressively built from fundamental physical properties. Importantly, ARSC successfully unified independent experimental data from 17 types of diatomic sites across 28 publications, with activity data exhibiting high consistency on the ARSC volcano plot, thereby validating the descriptor's universality and reliability.[77]

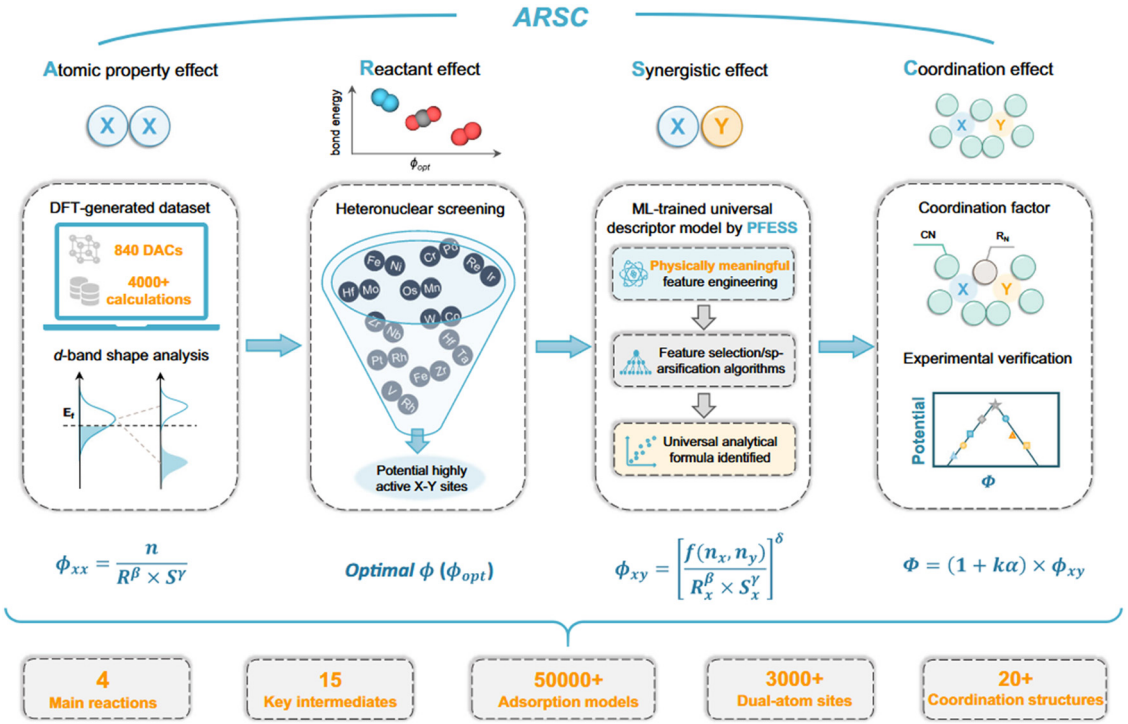

Fig. 5 General workflow of our work. Firstly, a primitive descriptor (φ_{xx}) for atomic property effects through d-band shape analysis. Secondly, screening principle (φ_{opt}) of potential desirable heteronuclear DACs based on reactant effects. Thirdly, ML-based descriptors (φ_{xy}) for synergistic effects through physically meaningful feature engineering based on φ_{xx} and feature selection/sparsification algorithms. Fourthly, the final universal descriptor model (Φ) with quantification of coordination effects and corresponding experimental verifications. Reproduced with permission.[77] Copyright 2024, Springer Nature.

SISSO has also shown exceptional capability in handling complex catalytic systems. Nair *et al.* developed an SISSO-guided active learning workflow, in which a closed-loop "predict-validate-update" mechanism enabled efficient screening of stable catalysts under acidic conditions.[33] To further enhance algorithmic practicality, the RF-SISSO model was introduced, achieving a 265-fold improvement in regression efficiency compared to the original SISSO model with only 45 samples.[78] Concurrently, manual feature engineering is increasingly integrated with interpretable ML; for instance, physically informed descriptors such as the "topological undercoordination number" have successfully revealed structural sensitivities in metal catalysts.[79]

These methodological refinements have substantially broadened the applicability of SISSO in machine-learning-assisted high-throughput screening of SACs across key electrocatalytic reactions. For example, high-throughput first-principles calculations combined with SISSO have been used to screen 192 transition-metal atoms anchored on 1T-TMD substrates for the CO_2RR, where SISSO-derived descriptors linking intrinsic features to limiting potentials guided the identification of promising catalysts such as Fe@CoS_2, Pt@$TiTe_2$, and Co@CoS_2 with low overpotentials and selective pathways to fuels like formic acid and methane.[80] In studies focusing on the HER, SISSO has been integrated into machine-learning workflows to derive interpretable descriptors correlating adsorption energetics and electronic structure with catalytic activity, enabling rapid screening and prediction of high-performance SACs.[81] Similarly, in ORR screening, SISSO-generated features such as combinations of d-electron count and Bader charge have been shown to play a critical role in predicting overpotentials and activity trends of MXene-supported single atoms.[82]

The SISSO method offers a robust and interpretable approach for small-sample datasets, automatically extracting optimal descriptors from large feature spaces while generating concise analytical expressions instead of black-box models. Its performance depends on the coverage of primary features, and large feature pools require pre-screening due to computational demands. Although its ability to capture highly nonlinear relationships is limited, SISSO serves as a powerful bridge between data-driven modeling and physical mechanism understanding, providing clear, actionable insights for rational design of heterogeneous catalysts.

3.2. Graph based features

3.2.1. Chemical graph features. While SISSO extracts interpretable descriptors to capture key physicochemical relationships, graph-based representations complement this approach by encoding atomic connectivity and local environments, providing ML models with structural information beyond what SISSO descriptors alone can offer. A chemical graph provides an

abstract yet effective framework for representing chemical systems, where atoms are modeled as nodes and chemical bonds or interactions as edges. Node features encode intrinsic atomic properties such as element type, electronegativity, valence electrons, and hybridization state, while edge features describe interatomic connections, including bond type, length, and order. This topological representation systematically encodes the chemical semantics of a system, enabling a unified framework for subsequent ML modeling. As one of the earliest and most mature feature representations, chemical graph descriptors illustrate how graph-based methodologies have evolved beyond conventional descriptors, laying the theoretical foundation for geometric graph features. Early applications in catalysis primarily targeted active site modeling and reaction network analysis. Notably, Dufaud *et al.* employed graph-theoretical concepts to design heterogeneous catalysts with multiple synergistic active sites, establishing a structured paradigm for modeling complex catalytic reactions.[83]

The development of chemical graph features originated in computational chemistry and chemoinformatics, where molecular systems were first abstracted as graphs with atoms as nodes and covalent bonds as edges. A notable milestone in this field was the establishment of the GDB-9 dataset by Ramakrishnan *et al.*, which systematically provided quantum chemical structural and property data for over 130 000 molecules. This dataset revealed the scaling behavior of chemical space with molecular size and the distribution of isomeric properties, validating the feasibility of predicting molecular properties through graph-based representations.[84] With subsequent advances, the concept of chemical graphs has been extended from molecular systems to crystalline and solid-state materials, supporting broad applications in materials science and heterogeneous catalysis. Concurrently, the advent of deep learning and GNNs has transformed chemical graph features from static descriptors into end-to-end, learnable representations. This evolution has culminated in the emergence of chemical graph neural networks, establishing a new paradigm for feature engineering in heterogeneous catalytic systems.

A breakthrough in chemical graph features for materials science and heterogeneous catalysis was achieved with the introduction of the CGCNN. In this framework, crystal structures are represented as graphs (Fig. 6a), where nodes correspond to atoms in the unit cell and edges denote chemical bonds or interatomic interactions. Graph convolutional layers iteratively update atomic features by aggregating information from neighboring atoms and bonds, while pooling layers integrate these local features into global crystal descriptors for property prediction. This end-to-end learning approach effectively eliminates the dependence on manually designed descriptors. Applied to ~47 000 crystal structures from the Materials Project, CGCNN achieved a mean absolute error of 0.039 eV per atom in formation energy prediction, surpassing conventional ML models.[85] Subsequently, Chen *et al.* developed the multi-task crystal graph convolutional neural network (MT-CGCNN), enabling efficient and accurate simultaneous prediction of multiple material properties. This multi-task framework is particularly advantageous in limited-data scenarios and high-throughput material screening.[86]

Researchers have leveraged the high efficiency of CGCNN in predicting energetic properties to enable high-throughput screening and prediction of catalysts.[25,87–90] Kim *et al.* introduced the Surface Graph Convolutional Neural Network (SGCNN), tailored to predict the binding energies of key adsorbates (*H, *N_2, *N_2H, *NH, and *N_2) relevant to the NRR. Using only low-dimensional inputs such as elemental properties and atomic connectivity, SGCNN achieved a mean absolute error of 0.23 eV on a dataset of 3040 DFT-calculated surfaces.[91] For HER catalysts, Zheng *et al.* employed an improved CGCNN model (ASB-GCNN) that partitions crystal geometry into active, surface, and bulk layers to screen 600 MA_2Z_4-based materials, identifying five promising SACs, including $V_1/HfSn_2N_4(S)$ with a near-ideal ΔG_{H^*} of 0.06 eV, thereby demonstrating efficient structure–activity mapping.[92] For OER catalysts, Back *et al.* combined DFT and CGCNN to identify low-index IrO_2 surfaces with lower overpotentials than the rutile(110) benchmark, highlighting GNN-assisted screening of active facets.[93] In the CO_2RR, Gu *et al.* used labeled site representations within a GNN framework to predict CO adsorption energies with an MAE of 0.116 eV, enabling rapid evaluation of diverse $Pd_xTi_{1-x}H_y$ surfaces.[94] Collectively, these developments highlight how graph-based neural networks have redefined chemical feature engineering, providing scalable and physically grounded representations for heterogeneous catalysis.

Leveraging chemical graph features, their application has been extended to automated construction and exploration of complex reaction networks. Zheng *et al.* combined graph theory with active learning to model the electro-synthesis of urea, representing reactants as molecular graphs (atoms as nodes, bonds as edges; Fig. 6b). Graph editing operations enabled automated simulation of elementary steps, constructing a reaction network with hundreds of intermediates. The graph stability prediction (GSP) algorithm reduced DFT computational cost by ~40% while maintaining accurate pathway identification (Fig. 6c and d).[95]

As model scale and catalytic system complexity increase, traditional GNNs face challenges in accurately capturing multiple adsorbates and diverse bond interactions. To address this, Bang *et al.* proposed the bond-type embedded crystal graph convolutional neural network (BE-CGCNN), which explicitly distinguishes and embeds four bond types: covalent, metallic, chemisorption, and nonbonded interactions, allowing more precise representation of nanoparticle surface chemistry (Fig. 7a and b). This approach discards distance-dependent features in favor of one-hot encoded bond types, enhancing robustness for unrelaxed structures and achieving a MAE of 0.07 eV in Pt *OH adsorption predictions (Fig. 7c).[96] For limited-data systems, Xu *et al.* developed a simplified crystal graph neural network with adaptive feature encoding (S-CGCNN), maintaining high predictive accuracy under small-sample conditions.[97] Chemical graph features are increasingly recognized as universal descriptors linking atomic structures to

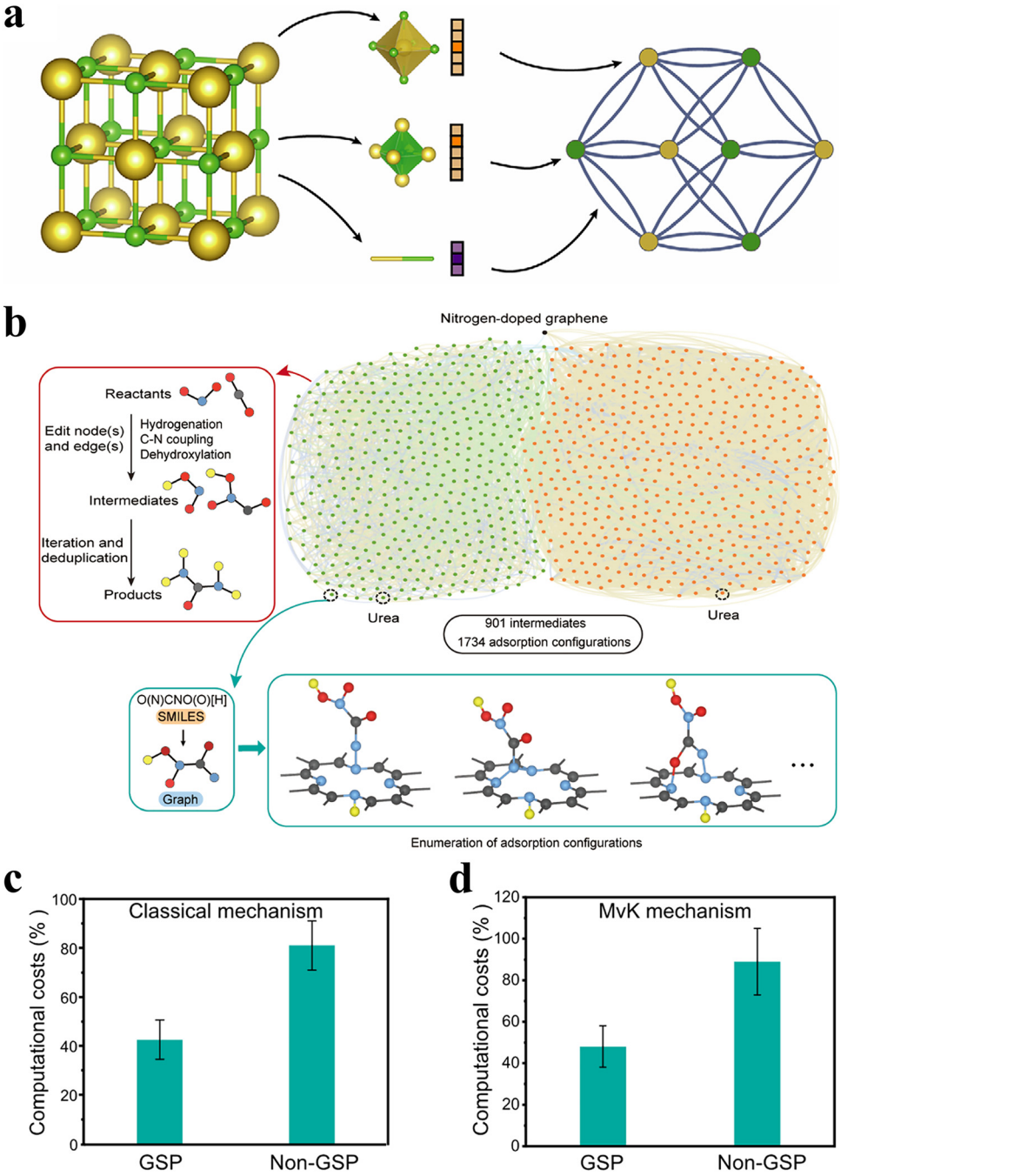

Fig. 6 (a) Construction of the reaction network for urea electrosynthesis on a nitrogen-doped carbon catalyst. Reproduced with permission.[56] Copyright 2018, American Physical Society. (b) A scheme of the overall reaction network, the typical elementary steps, and the example of adsorption configuration enumeration. Green dots represent reaction intermediates in the classical mechanism, while orange dots represent those in the MvK mechanism. The average DFT calculation costs (percentage of reaction intermediates calculated by DFT) for the prediction of the (c) classical mechanism and (d) MvK mechanism with and without the GSP algorithm. Reproduced with permission.[95] Copyright 2025, Chinese Chemical Society.

macroscopic properties, providing a foundation for high-throughput screening, multi-task learning, and active site mechanistic analysis.[98,99]

3.2.2. Geometric graph features. Building on chemical graph features, geometric graphs incorporate three-dimensional atomic coordinates and spatial relationships to capture complex catalytic environments. Nodes encode atomic properties, edges describe interatomic distances and angles, and global features reflect lattice parameters and symmetry. Unlike molecular graphs based solely on chemical bonds, geometric graphs employ cutoff radii or Voronoi analysis to capture long-range interactions and steric effects, enabling accurate representation of adsorption, coordination, and electronic structure on catalyst surfaces.

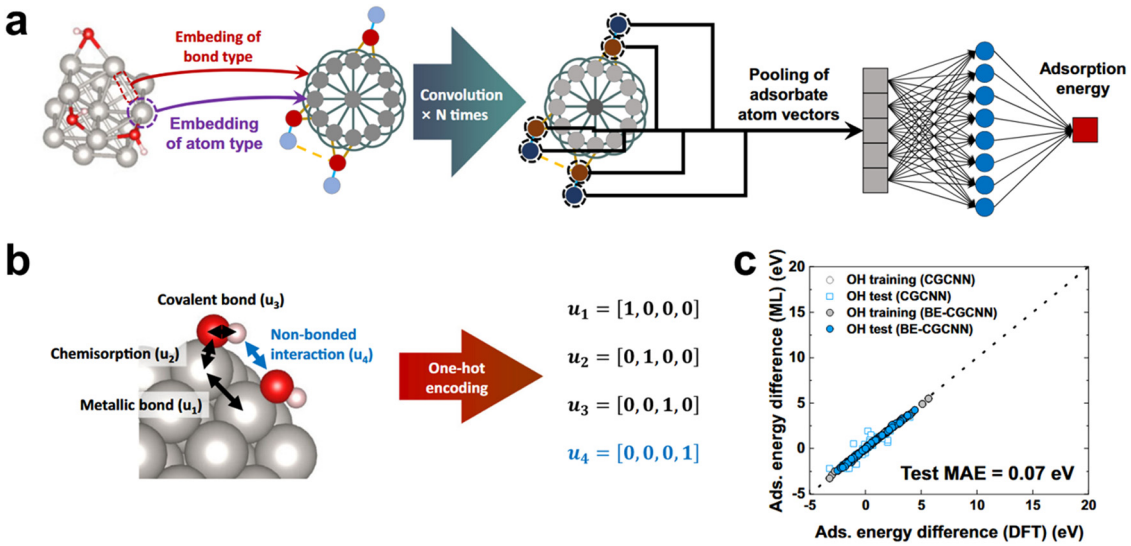

Fig. 7 (a) A schematic representation of the graph convolution neural network model to predict the adsorption energy; (b) representation of bond embedding. Each bond is embedded into a bond vector by one-hot encoding of the bond type; (c) comparison of BE-CGCNN and CGCNN predictions for adsorption energy differences on the OH adsorbate dataset. Reproduced with permission.[96] Copyright 2023, Springer Nature.

SchNet, introduced by Schütt *et al.*, pioneered the use of continuous-filter convolution to model quantum interactions in molecular and crystalline systems, enabling direct learning of potential energy surfaces from atomic 3D coordinates. Unlike traditional graph-based chemical descriptors, SchNet represents atoms as nodes and spatial vectors between atoms (distance and direction) as edge features, encoded continuously *via* radial basis functions. As illustrated in Fig. 8a, a learnable filter network processes this geometric information through multiple interaction blocks, progressively updating atomic features to output rotationally invariant total energies and rotationally covariant atomic forces.[100] Building on this approach, Chen *et al.* developed the MEGNet (Materials Graph Network) framework, which unifies the treatment of molecules and crystals within a graph neural network by incorporating multi-level updates for atoms, bonds, and global state variables (*e.g.*, temperature, pressure) (Fig. 8b). Trained on ~69 000 crystals from the Materials Project, MEGNet outperformed prior models such as CGCNN in predicting crystal formation energies and bulk moduli.[101]

The DimeNet model (Fig. 9a) overcomes the limitations of distance-only geometric representations by introducing directional message passing. Using a two-dimensional basis of spherical Bessel and spherical harmonic functions, it explicitly encodes both distances and bond angles, enabling precise modeling of local directional interactions.[102] This approach is particularly effective for analyzing conformational evolution in multi-step reactions such as the OER and CO_2RR. Building on this, Li *et al.* proposed LEPool-DimeNet++, which incorporates local environment pooling to improve adsorption energy predictions. The model achieved mean absolute errors (MAEs) of 0.096 eV and 0.073 eV for *CO and *H adsorption energies, respectively, outperforming previous state-of-the-art models.[104] Further advancing geometric graph representations, Choudhary *et al.* developed ALIGNN (Atomistic Line Graph Neural Network) (Fig. 9b). ALIGNN performs message passing simultaneously on atomic graphs (nodes = atoms, edges = bonds) and line graphs (nodes = bonds, edges = bond angles), collaboratively updating atomic, bond, and bond-angle features. This allows accurate capture of local geometric configurations at surface active sites and shows exceptional performance in adsorption energy prediction, conformational stability analysis, and modeling reaction intermediates.[103] Geometric graph neural networks also extend to complex organic molecules. GAME-Net, developed by Pablo-García *et al.* (Fig. 10), predicts adsorption energies of organic molecules on metal surfaces with an MAE of 0.18 eV, achieving approximately six orders of magnitude faster computation than traditional DFT, thereby enabling high-throughput screening of heterogeneous catalysts.[26]

Geometric graph features naturally encode three-dimensional structural information. This makes them well-suited for complex catalytic systems, including alloys, surfaces, and interfaces. Nevertheless, these features face several challenges. They are sensitive to unrelaxed structures and highly dependent on precise atomic positions, often requiring large training datasets for complex systems. Their high computational complexity also places greater demands on algorithms and hardware.

3.3. Topological data features

While geometric features capture atomic spatial arrangements, they are sensitive to structural relaxation and local perturbations. To extract descriptors robust to such deformations, algebraic topology methods have been introduced. These approaches characterize the intrinsic connectivity of materials by identifying discrete entities, ring-like structures, and high-dimensional cavities *via* computing topological invariants that

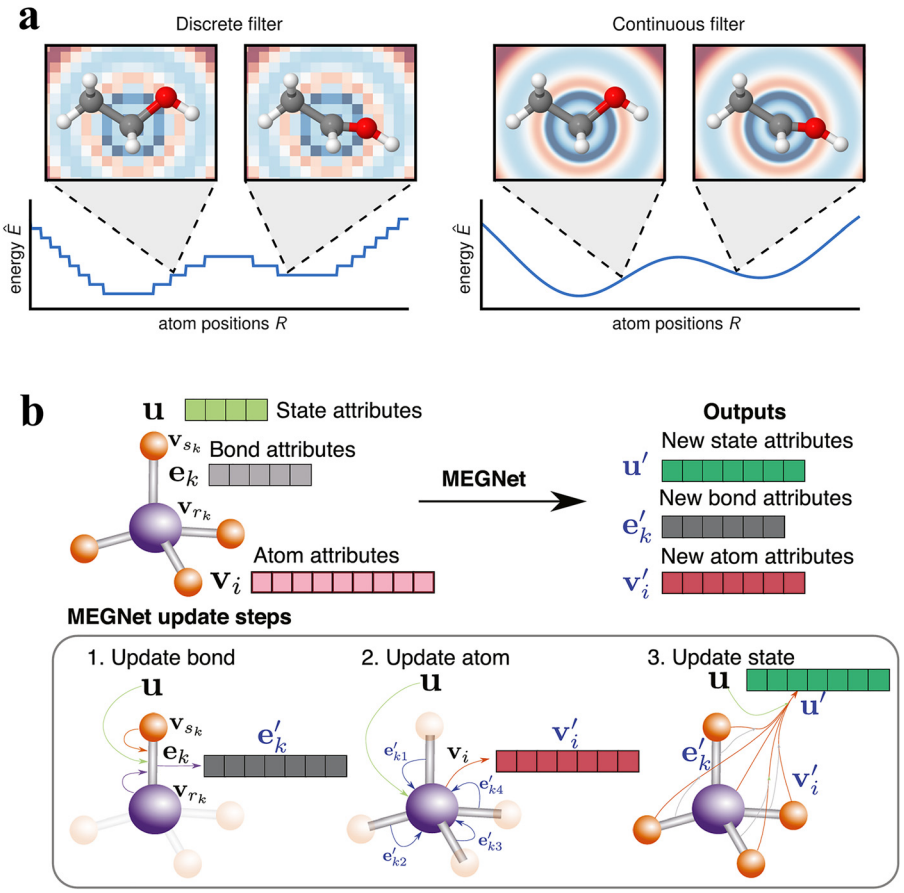

Fig. 8 (a) The discrete filter (left) is not able to capture the subtle positional changes of the atoms, resulting in discontinuous energy predictions E (bottom left). The continuous filter captures these changes and yields smooth energy predictions (bottom right). Reproduced with permission.[100] Copyright 2017, the Authors. (b) A MEGNet module starts with atomic attributes $V = \{v_i\}_{i=1:N}^{v}$, and $E = \{(e_k, r_k, s_k)\}_{k=1:N}^{e}$ global state attributes. Through sequential updates of bonds, atoms, and the global state, information flows among all three, yielding a new graph representation. Reproduced with permission.[101] Copyright 2019, American Chemical Society.

remain unchanged under continuous deformations.[105] Homology groups, a core concept in algebraic topology, detect "holes" of different dimensions through simplicial complexes and boundary operators. The number of such features is quantified by Betti numbers (β_0, β_1, β_2...), representing connected components, one-dimensional loops, and two-dimensional cavities.[106,107] Traditional homology is static and does not capture scale-dependent features. Persistent homology overcomes this limitation by constructing nested topological spaces through increasing scale parameters (*e.g.*, interatomic distances or electron densities), tracing the birth and death of topological data features.[108–111] These results are visualized as persistent barcodes or diagrams, with bar length reflecting feature stability (Fig. 11a).[108] Topological descriptors capture both local atomic arrangements and global structural patterns, including long-range interactions. They have demonstrated success in bioactivity prediction, protein structure analysis, membrane fusion pore studies, and gene regulatory networks,[108,110,112–116] and are increasingly applied to materials science and heterogeneous catalysis, highlighting their growing relevance in complex materials modeling.[117,118]

In heterogeneous catalysis, adsorption energies at active sites critically determine catalyst activity, selectivity, and stability. These energies strongly depend on local atomic configurations, coordination environments, and electronic structures, making the establishment of universal structure–performance relationships challenging. Topological data features provide a machine-learning-compatible representation that captures complex three-dimensional structures. For example, in metal–organic frameworks (MOFs), pore topology governs gas adsorption behavior. Yang *et al.* applied TDA to convert MOF crystal structures into descriptors quantifying pore connectivity, ring structures, and cavity distribution (Fig. 11b). When combined with an extreme gradient boosting (XGBoost) model, this approach substantially outperformed traditional geometric descriptors in predicting C_1–C_3 alkane adsorption performance.[117]

TDA can capture structural information that is difficult to encode using conventional graph-based representations. For example, in metal–nitrogen–carbon SACs, local curvature plays a critical role in modulating the geometric environment. However, such curvature does not directly alter bonding connectivity and therefore remains challenging to represent using

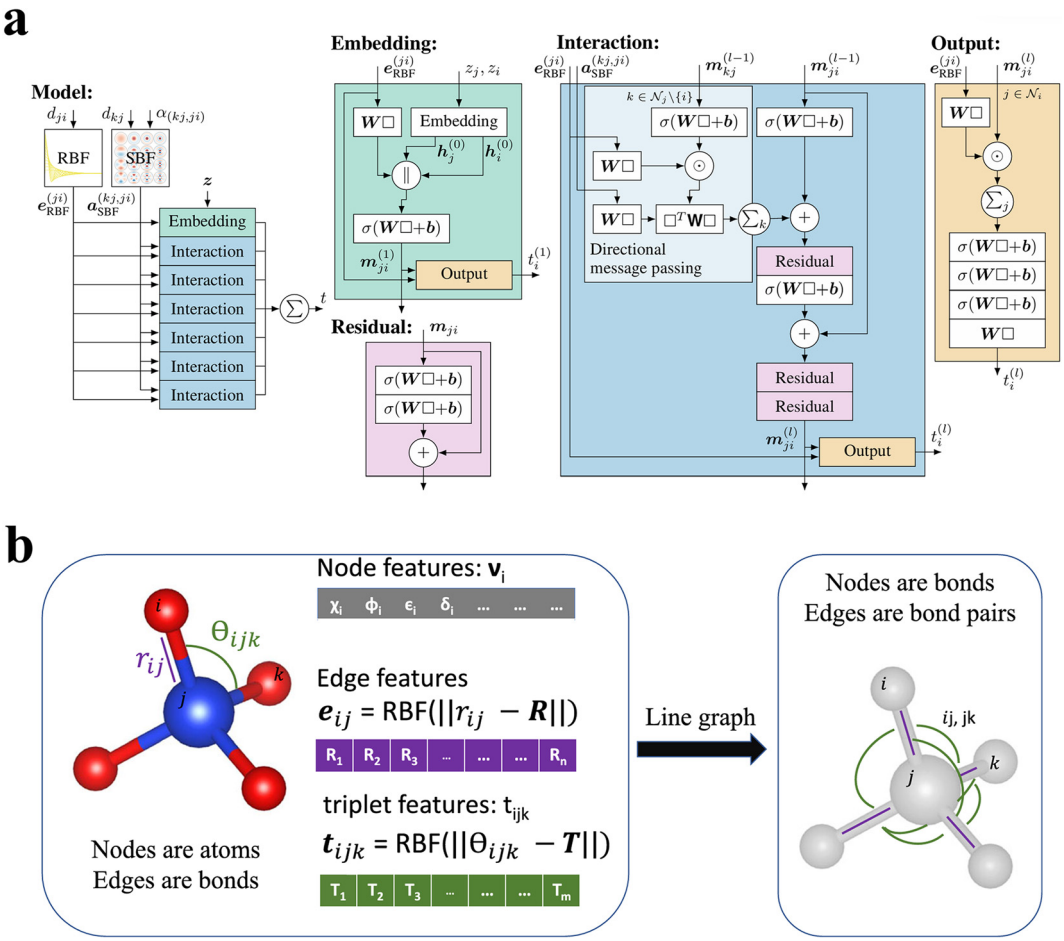

Fig. 9 (a) The DimeNet architecture represents distances (d_{ji}) using spherical Bessel functions and distances (d_{kj}) with angles ($a_{kj,ji}$) using a 2D spherical Fourier–Bessel basis. An embedding block generates initial message embeddings (m_{ji}), which are updated through multiple interaction blocks *via* directional message passing using neighboring messages (m_{kj}), 2D representations ($a_{kj,ji}^{SBF}$), and distance representations (e_{ji}^{RBF}). Each block outputs transformed embeddings using (e_{ji}^{RBF}) and sums them per atom, and the outputs of all layers are finally summed to yield the prediction. Reproduced with permission.[102] Copyright 2022, the Authors. (b) An undirected crystal graph representation and the corresponding line graph construction of a SiO_4 polyhedron. For clarity, only Si–O bonds are shown. The ALIGNN convolution layer alternates messages passing between the bond graph (left) and the line graph (bond adjacency graph, right). Reproduced with permission.[103] Copyright 2021, Springer Nature.

standard graph encodings. Liang *et al.* developed the persistent homology-enhanced crystal graph convolutional neural network (PH-CGCNN) (Fig. 12), which embeds curvature-induced microstructural variations by persistent homology into graph neural network features. The barcodes generated by persistent homology can effectively distinguish structural variations induced by different curvature conditions.

Beyond interpretation and prediction, topological approaches have been extended to reverse design of active sites. Wang *et al.* developed a topology-based variational autoencoder framework (PGH-VAEs) that represents catalytic sites using persistent GLMY homotopic features (Fig. 13a–c). This method quantifies the relationship between three-dimensional structural sensitivity and adsorption properties, enabling interpretable design of high-entropy alloy active sites. Applied to the IrPdPRhRu system, it revealed synergistic regulation of *OH adsorption energies by coordination and ligand effects. Latent-space analysis identified Pt–Pd bridging sites combined with distal Ru atoms as optimal configurations, exhibiting higher OH adsorption energies than Pt(111). Incorporating second-neighbor Ru further modulated the d-band center, optimizing the ORR pathway and enhancing catalytic efficiency and poisoning resistance.[120]

Zheng *et al.* proposed PH-SA (Fig. 14a–d) to efficiently explore active phase configurations using TDA. The method decomposes structures into atomic aggregates, identifies potential adsorption sites *via* persistent homotopy, and generates configurations through combinatorial enumeration. Machine learning force fields optimize these structures, and Pourbaix diagrams track phase evolution under external conditions. PH-SA samples surface, subsurface, and bulk sites for both slabs and clusters, overcoming limitations of intuition-based approaches. In Pd hydrogenation and Pt cluster oxidation, it accurately predicted structural rearrangements and reactivity, providing an efficient framework for discovering active phases and elucidating catalytic mechanisms.[121]

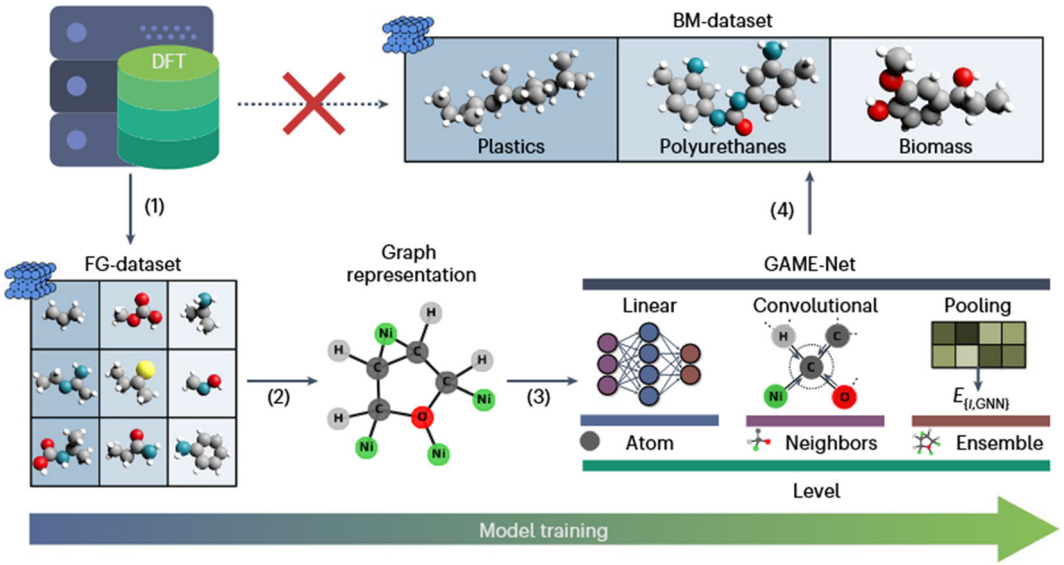

Fig. 10 Schematic illustration of the GAME-Net workflow. Starting from the DFT FG-dataset of small adsorbates (3315 samples), adsorption systems are converted into graph representations to train the proposed GNN architecture. GAME-Net is then applied to predict adsorption energies of larger molecules ($C < 23$) on metal surfaces in the BM-dataset, eliminating the need for costly DFT calculations. Reproduced with permission.[26] Copyright 2023, Springer Nature.

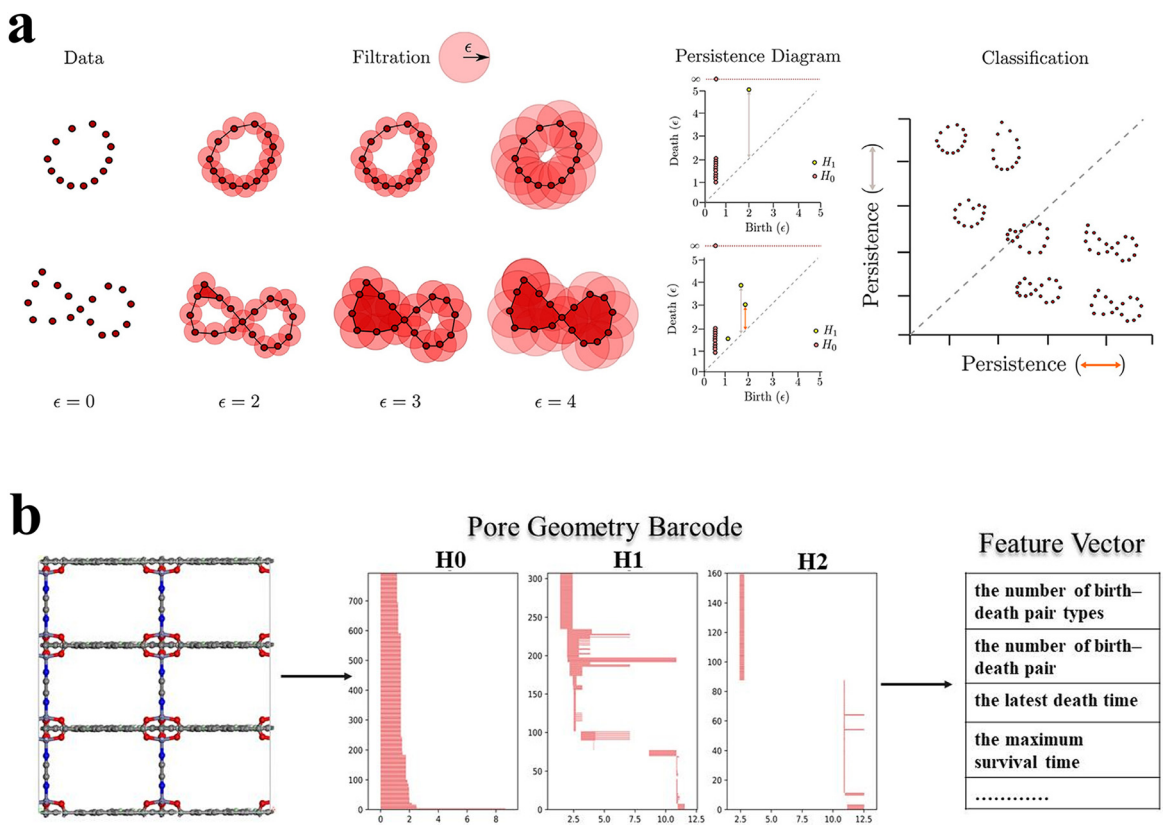

Fig. 11 (a) Schematic illustration of the persistent homology methodology for point clouds. Each point cloud is transformed into a geometric object through filtration, during which topological data features (*e.g.*, holes) appear and disappear. Their birth and death values are recorded in a persistence diagram (x = birth, y = death, persistence = $y - x$), capturing the topological evolution of the data. The persistence diagram enables distinguishing point clouds of different shapes and clustering those with similar topology, as demonstrated by the representative classification plot. Reproduced with permission.[108] Copyright 2021, Elsevier. (b) The flowchart illustrates the development of the topological descriptor for hMOF5035530. In the pore geometry barcode, the horizontal axis represents the filter radius, while the vertical axis indicates the number of barcodes. Reproduced with permission.[117] Copyright 2024, MDPI.

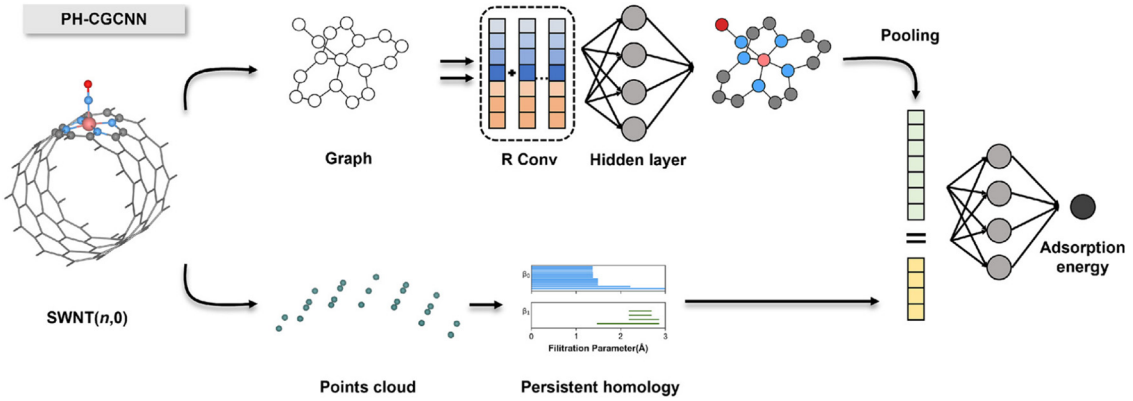

Fig. 12 Architecture of the PH-CGCNN model, combining atomic graph representations from standard CGCNN with persistent homology-derived curvature features to predict adsorption energies. Reproduced with permission.[119] Copyright 2025, Elsevier.

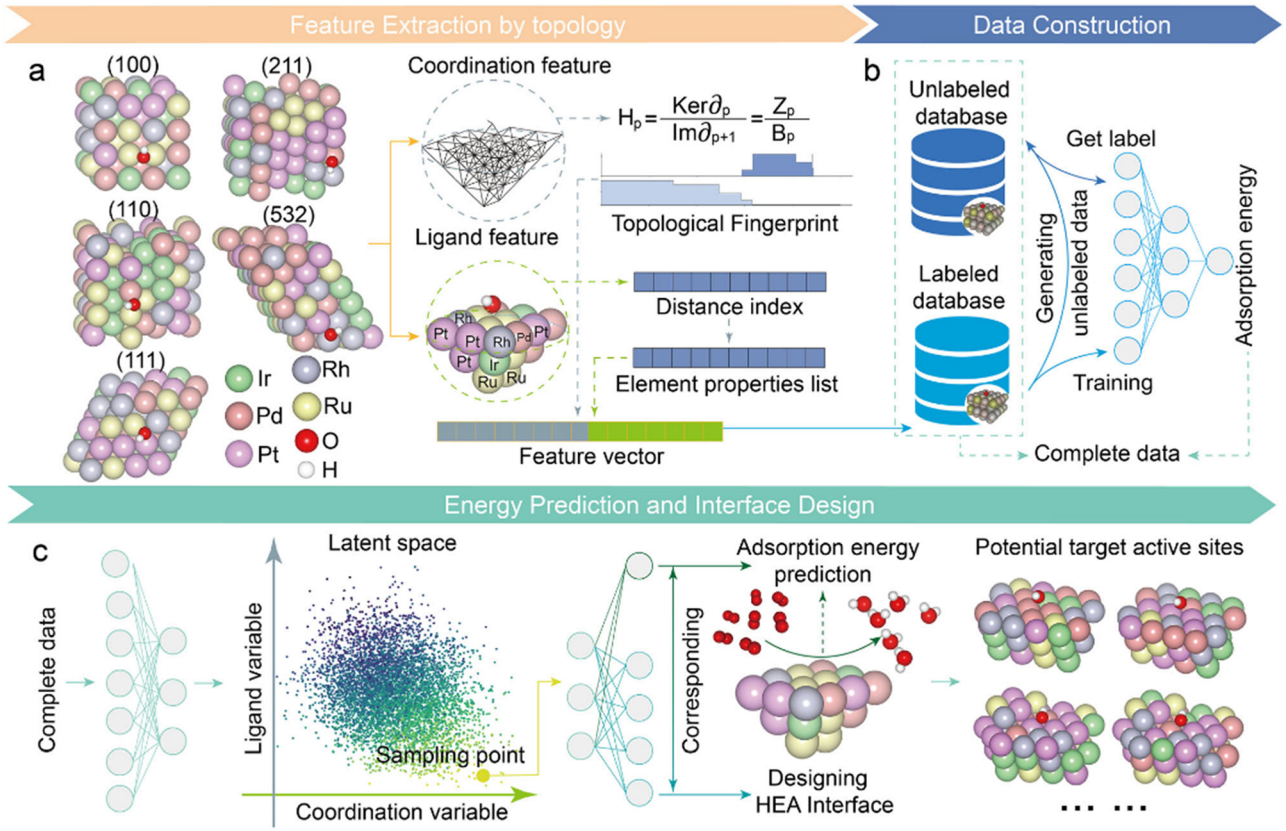

Fig. 13 (a) Schematic of feature construction, where coordination features are obtained *via* the PGH method and ligand features are represented using elemental properties. (b) Dataset construction using DFT and semi-supervised learning: a GBR model is first trained on DFT-calculated adsorption energies and then used to predict energies for additional simulated active sites, creating an expanded pseudo-labeled dataset for model training. (c) Framework of PGH-VAEs, showing modules for encoding, latent space visualization, sampling, and decoding to generate potential active sites. Reproduced with permission.[120] Copyright 2025, Chinese Academy of Sciences.

Topological algebra methods provide multiscale representations for designing heterogeneous catalysts and predicting performance. Challenges include integrating element-specific information and combining topological descriptors with electronic structure features. Techniques such as attention mechanisms, multimodal learning, and dynamic descriptors can enhance expressiveness and applicability. With further development, they hold promise for elucidating catalytic mechanisms, guiding material design, and accelerating catalyst discovery.

3.4. Multimodal features

Single-descriptor models often fail to capture the full complexity of heterogeneous catalysts. Multimodal feature fusion integrates

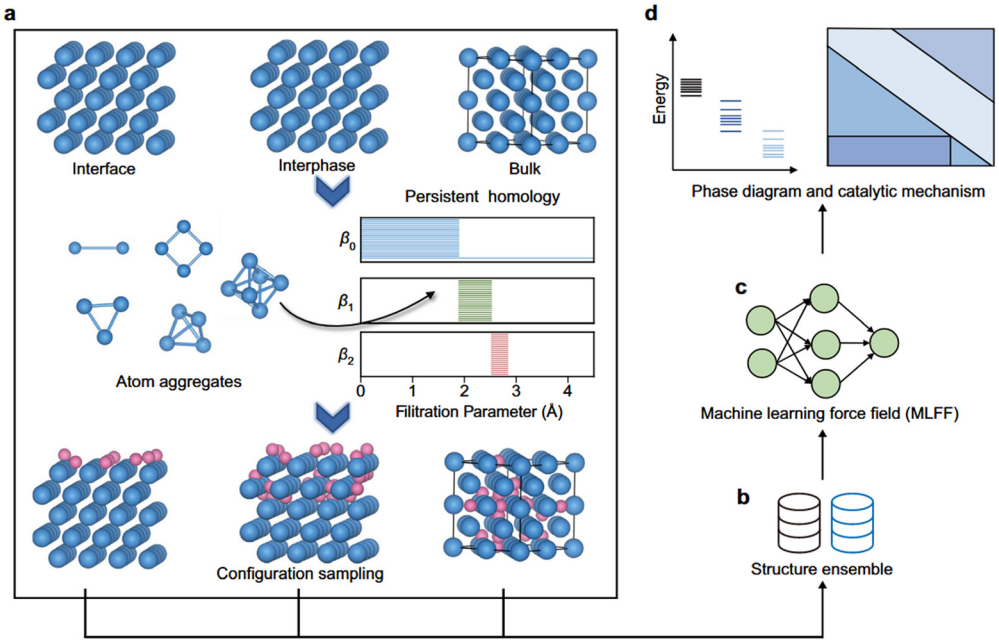

Fig. 14 (a) The PH-SA decomposes material structures into small atomic aggregates, using persistent homology to identify potential interaction sites within each unit. Combining the sites from all aggregates yields the potential active sites for species across the entire structure. (b) Identified sites are used in combinatorial enumeration to generate a set of structures. (c) A machine learning force field (MLFF) is trained *via* transfer learning to improve computational efficiency. (d) The Pourbaix diagram under specific conditions is constructed to aid catalytic mechanism analysis. Reproduced with permission.[121] Copyright 2025, Springer Nature.

computational and experimental data, including composition, atomic structure, electronic properties, and macroscopic performance, into high-dimensional "super-descriptors", enabling a more comprehensive, mechanism-driven understanding of catalytic behavior.[122] For instance, the adsorption multi-modal transformer model proposed by Chen *et al.* effectively couples atomic-level graph representations of catalyst surfaces with molecular descriptors of adsorbates *via* a cross-attention mechanism, thereby achieving high-precision predictions of global minimum adsorption energies.[9]

Textual representations can complement traditional descriptors, supporting multimodal approaches for catalyst design. Catalyst generative pretrained transformer (CatGPT) generates chemically valid string representations of inorganic catalysts from text inputs. When fine-tuned on specialized datasets, such as binary alloys for the two-electron ORR, it can propose candidate structures tailored to specific catalytic applications.[123] As a specific implementation of multimodal fusion, structure-text alignment aims to connect textual material descriptions with numerical structural features. Ock *et al.* developed a graph-assisted pretraining framework that integrates atomic structures with textual descriptions for predicting adsorption energies. As shown in Fig. 15a, the framework operates through two main stages: self-supervised graph-text alignment pretraining followed by supervised fine-tuning for energy prediction. This approach aligns graph embeddings from an equivariant graph neural network (EquiformerV2) with text embeddings from a Transformer language model (CatBERTa) in a shared latent space (Fig. 15b). The framework also incorporates a large language model (CrystalLLM) to generate structural descriptions from simplified chemical inputs, enabling reasonable energy predictions without complete atomic coordinates (Fig. 15c). This method offers a potential pathway for utilizing textual information from the literature to support catalyst screening.[21]

In multimodal catalyst screening, spectroscopic descriptors complement conventional structural representations and enable integrated AI-driven design. To improve catalyst screening efficiency, Yang *et al.* developed a cross-modal encoder–decoder framework that integrates spectroscopic and structural descriptors for comprehensive chemical representation. Using pretraining strategies based on property regression and masked-mode prediction, the model enables bidirectional translation between molecular geometries and vibrational spectra. In CO/NO adsorption on Ag/Au surfaces, the framework exploits the complementarity of infrared and Raman signals to accurately predict adsorption properties and internal coordinates (RMSE $\approx$ 0.01 Å). This method addresses the information insufficiency of single-modality representations and supports multi-objective prediction and data recovery in complex catalyst design.[125] Subsequently, Zhao *et al.* extended multimodal fusion to organic molecular structure elucidation by proposing a framework based on one-dimensional convolutional neural networks (1D-CNNs). By integrating infrared, Raman, and nuclear magnetic resonance spectra, the method effectively leverages the complementary strengths of vibrational and magnetic resonance information, enabling automated identification and quantification of functional groups.[126] Although several multimodal alignment

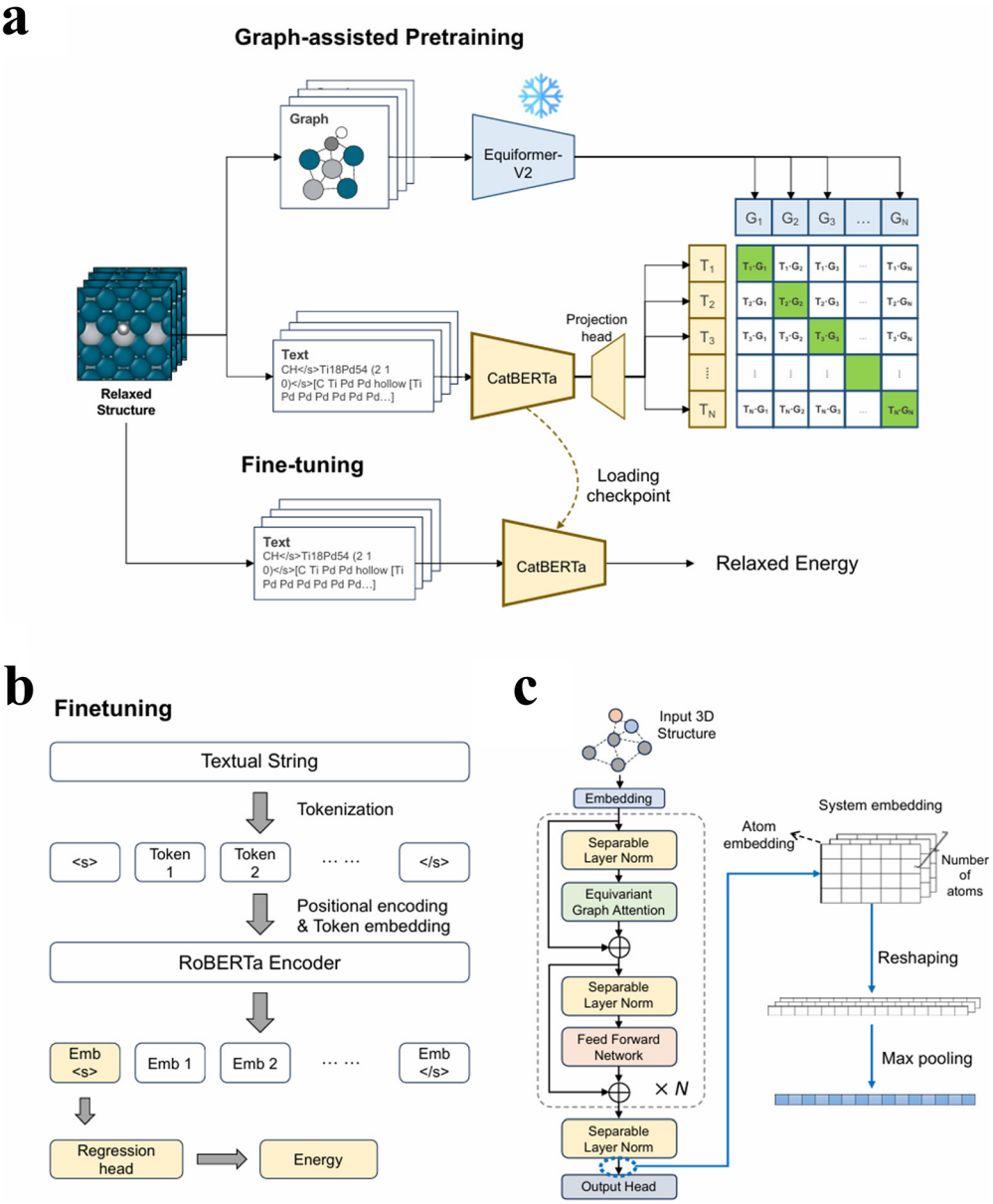

Fig. 15 (a) The training involves two steps: graph-assisted pretraining followed by energy prediction fine-tuning. (b) The CatBERTa model is used as the text encoder. (c) The graph encoder, with the final-layer graph embeddings reshaped and max-pooled into a 1D format. The architecture is reproduced from the original EquiformerV2[124] publication. Reproduced with permission.[21] Copyright 2025, Springer Nature.

approaches integrating spectra, structures, and text have emerged in recent years, such as TranSpec and SpecGNN, which enable bidirectional translation between vibrational spectra and SMILES representations, their application in catalysis remains relatively unexplored.[127]

Compared to traditional single-modality descriptors, multimodal feature representations offer a more comprehensive capture of the complex structure–property relationships in catalytic systems. As a result, they have emerged as a critical direction in structural feature engineering, showing great potential in enhancing both the generalization ability and physical interpretability of ML models.

4. Summary and outlooks

Structural feature engineering in heterogeneous catalysis has made significant progress, particularly in establishing generalizable models across diverse material systems. GNNs, such as DimeNet and GAMENet, have emerged as powerful tools capable of capturing atomic connectivity and local environments, and have been widely applied to structure–property predictions in metals, oxides, and alloy-based catalytic systems. In terms of interpretability, the introduction of sparse modeling approaches like SISSO has enabled the automatic identification of key atomic-scale descriptors directly correlated with catalytic activity, offering physically meaningful insights into

structure–performance relationships. Despite this progress, structural feature engineering in heterogeneous catalysis still faces numerous challenges, necessitating sustained and interdisciplinary collaboration.

4.1. The in-depth development of multimodal features

Multimodal feature integration for structural representation in heterogeneous catalysis is still at an early stage. Most existing studies have focused on limited modality combinations, such as structure–text[21] and structure–spectra integration.[128] Within this scope, structure–Raman fusion represents a typical form of structure–spectroscopy multimodality, where vibrational fingerprints provide chemically informative, albeit indirect, insights into local bonding environments and surface intermediates. However, extending these approaches to higher-dimensional experimental modalities, including charge density maps, electron microscopy images (*e.g.*, STEM and TEM), and spectroscopic signals such as XAS and XRD, remains challenging. These modalities differ fundamentally from atomistic structural descriptors in data format, spatial resolution, and physical semantics, hindering their direct incorporation into unified machine-learning frameworks. The key challenge lies in developing representations that simultaneously maintain structural sensitivity and physical interpretability while enabling automated extraction of catalytically relevant features, such as defects, coordination environments, and local electronic states. Despite the progress detailed in Section 3.4, practical multimodal applications in catalysis remain scarce. The field is pivoting toward unified multimodal learning architectures that transcend simple feature concatenation. By employing cross-attention or contrastive learning, these frameworks deeply align atomic topologies, electronic states, and experimental fingerprints within a shared latent space.[129–131]

4.2. Interpretability enhancement

Although notable advances have been achieved in enhancing interpretability for machine-learning–assisted catalytic design, key limitations remain. Sparse symbolic regression methods such as SISSO yield low-dimensional, physically interpretable descriptors that enable mechanistic insight,[74] but their reliance on predefined operator sets and sparse formulations restricts their capacity to capture strong nonlinearities, high-order interactions, and complex reaction dynamics.[72] In contrast, deep learning models, particularly graph neural networks, provide higher predictive accuracy and generalization by modeling intricate structure–property relationships, yet their limited interpretability hampers physical understanding and hypothesis generation.[132] To bridge the gap between predictive power and interpretability in deep learning, two promising directions have emerged. The first involves hybrid frameworks that integrate GNNs with sparse modeling approaches like SISSO, where the rich structural representations learned by GNNs serve as input features for interpretable models, achieving a balance between expressiveness and transparency. The second path focuses on developing inherently interpretable deep learning architectures. Among these, topological deep learning shows great promise.[119,120]

4.3. Cross-scale structural mapping

Catalyst structures inherently span multiple spatial scales, from atomic coordination environments and point defects to surface reconstructions, crystal facets, particle morphology, and macroscopic catalytic architectures, posing a fundamental challenge for structure–property modeling. While atomistic simulations such as DFT are effective in identifying active sites and elementary mechanisms, they are intrinsically limited in capturing mesoscale and macroscale effects, whereas experimentally or mesoscale-derived structural information often lacks atomistic resolution. Most current machine-learning representations therefore remain confined to microscopic atomic configurations, with insufficient treatment of larger-scale structural features. Beyond spatial scale mismatch, catalytic modeling also suffers from pronounced fidelity inconsistency arising from different levels of physical approximation, ranging from low-cost DFT calculations that neglect solvation or electrochemical potential, to high-fidelity constant-potential simulations and experimental conditions. Multi-fidelity learning, particularly Δ-learning, offers a systematic strategy to address this issue by learning structure-preserving corrections that map low-fidelity predictions onto higher-level physical descriptions, thereby reconciling computational efficiency with physical accuracy.[133,134] Although such approaches have been actively explored in materials science and molecular modeling,[135,136] their application in heterogeneous and electrocatalysis remains limited, highlighting a critical opportunity for catalytic research. Integrating multi-scale structural representations with multi-fidelity learning frameworks is therefore expected to enable consistent structure–property relationships across spatial hierarchies and levels of theory, bridging electronic-structure simulations, realistic operating conditions, and device-level catalytic performance toward unified, system-level catalyst design.

In summary, this review systematically summarizes advances in feature engineering for heterogeneous catalysis, from empirical descriptors to data-driven methods. Early work relied on intuitive descriptors like the atomic radius and electronegativity, linking electronic structure to activity but struggling with high-dimensional complexity. Symbolic regression methods, such as SISSO, automate the discovery of low-dimensional, interpretable descriptors, improving prediction and mechanistic insight. GNNs model catalysts as atomistic graphs, capturing local environments and multi-site interactions, enhancing adsorption and activity predictions. TDA, *e.g.*, persistent homology, provides multi-scale insights into structural connectivity. Integrating multimodal data from computational, experimental, and literature sources enables mechanism-informed ''super-descriptors,'' bridging prediction and understanding. Overall, feature engineering advances both predictive accuracy and mechanistic insight, supporting rational, data-driven design of next-generation heterogeneous catalysis.

Author contributions

Yu Jin: conceptualization, original draft, review & editing. Hang-Biao Lv: original draft. Shisheng Zheng: conceptualization, review & editing, project administration, supervision. Jian-Feng Li: supervision, funding acquisition.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (22402163, 22021001, 21925404, T2293692, and 22361132532), the Natural Science Foundation of Xiamen, China (3502Z202472001), the Fundamental Research Funds for the Central Universities (20720250026) and the Fujian Provincial Science and Technology Program for International Cooperation (2025I0002).

References

1 Y. Zang, D.-Q. Lu, K. Wang, B. Li, P. Peng, Y.-Q. Lan and S.-Q. Zang, A pyrolysis-free Ni/Fe bimetallic electrocatalyst for overall water splitting, *Nat. Commun.*, 2023, **14**(1), 1792.
2 L. Yan, P. Li, Q. Zhu, A. Kumar, K. Sun, S. Tian and X. Sun, Atomically precise electrocatalysts for oxygen reduction reaction, *Chem*, 2023, **9**(2), 280–342.
3 Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov and T. F. Jaramillo, Combining theory and experiment in electrocatalysis: Insights into materials design, *Science*, 2017, **355**(6321), eaad4998.
4 N. S. Muhammed, B. Haq, D. Al Shehri, A. Al-Ahmed, M. M. Rahman and E. Zaman, A review on underground hydrogen storage: Insight into geological sites, influencing factors and future outlook, *Energy Rep.*, 2022, **8**, 461–499.
5 M. N. Hossain, L. Zhang, R. Neagu and E. Rassachack, Free-standing single-atom catalyst-based electrodes for CO2 reduction, *Electrochem. Energy Rev.*, 2024, **7**(1), 5.
6 H. Ding, S. Zheng, X. Yang, J. Pan, Z. Chen, M. Zhang, S. Li and F. Pan, Role of surface hydrogen coverage in C–C coupling process for CO2 electroreduction on ni-based catalysts, *ACS Catal.*, 2024, **14**(19), 14330–14338.
7 T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-I. Shimizu, Machine learning for catalysis informatics: recent applications and prospects, *ACS Catal.*, 2019, **10**(3), 2260–2297.
8 L. Zhang, Q. Bing, H. Qin, L. Yu, H. Li and D. Deng, Artificial intelligence for catalyst design and synthesis, *Matter*, 2025, **8**(5), 102138.
9 J. Chen, X. Huang, C. Hua, Y. He and P. Schwaller, A multimodal transformer for predicting global minimum adsorption energy, *Nat. Commun.*, 2025, **16**(1), 3232.
10 Z. Fu, P. Huang, X. Wang, W. D. Liu, L. Kong, K. Chen, J. Li and Y. Chen, Artificial Intelligence-Assisted Ultrafast High-Throughput Screening of High-Entropy Hydrogen Evolution Reaction Catalysts, *Adv. Energy Mater.*, 2025, **15**(30), 2500744.
11 Z. Zhang, S. Ma, S. Zheng, Z. Nie, B. Wang, K. Lei, S. Li and F. Pan, Semantic knowledge graph as a companion for catalyst recommendation, *Nat. Sci. Open*, 2024, **3**(2), 20230040.
12 X. Jia, T. Wang, D. Zhang, X. Wang, H. Liu, L. Zhang and H. Li, Advancing electrocatalyst discovery through the lens of data science: State of the art and perspectives, *J. Catal.*, 2025, **447**, 116162.
13 J. Yin, H. Chen, J. Qiu, W. Li, P. He, J. Li, I. A. Karimi, X. Lan, T. Wang and X. Wang, SurFF: a foundation model for surface exposure and morphology across intermetallic crystals, *Nat. Comput. Sci.*, 2025, **5**, 782–792.
14 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**(7715), 547–555.
15 L. H. Mou, T. Han, P. E. Smith, E. Sharman and J. Jiang, Machine learning descriptors for data-driven catalysis study, *Adv. Sci.*, 2023, **10**(22), 2301020.
16 S. N. Steinmann, Q. Wang and Z. W. Seh, How machine learning can accelerate electrocatalysis discovery and optimization, *Mater. Horiz.*, 2023, **10**(2), 393–406.
17 Z. Li, C. Zhao, H. Wang, Y. Ding, Y. Chen, P. Schwaller, K. Yang, C. Hua and Y. He, Interpreting chemisorption strength with AutoML-based feature deletion experiments, *Proc. Nat. Acad. Sci. U. S. A.*, 2024, **121**(12), e2320232121.
18 J. G. T. Tomacruz, K. E. S. Pilario, M. F. M. Remolona, A. A. B. Padama and J. D. Ocon, A machine learning-accelerated density functional theory (ML-DFT) approach for predicting atomic adsorption energies on monometallic transition metal surfaces for electrocatalyst screening, *Chem. Eng. Trans.*, 2022, **94**, 733–738.
19 A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock and T. D. Sparks, Compositionally restricted attention-based network for materials property predictions, *npj Comput. Mater.*, 2021, **7**(1), 77.
20 J. Lan, A. Palizhati, M. Shuaibi, B. M. Wood, B. Wander, A. Das, M. Uyttendaele, C. L. Zitnick and Z. W. Ulissi, AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials, *npj Comput. Mater.*, 2023, **9**(1), 172.
21 J. Ock, S. Badrinarayanan, R. Magar, A. Antony and A. Barati Farimani, Multimodal language and graph learning of adsorption configuration in catalysis, *Nat. Mach. Intell.*, 2024, **6**(12), 1501–1511.
22 M. Weng, Z. Wang, G. Qian, Y. Ye, Z. Chen, X. Chen, S. Zheng and F. Pan, Identify crystal structures by a new

paradigm based on graph theory for building materials big data, *Sci. China: Chem.*, 2019, **62**(8), 982–986.
23 X. Wan, Z. Zhang, W. Yu, H. Niu, X. Wang and Y. Guo, Machine-learning-assisted discovery of highly efficient high-entropy alloy catalysts for the oxygen reduction reaction, *Patterns*, 2022, **3**(9), 100553.
24 T. Vinchurkar; J. Ock and A. B. Farimani Explainable Data-driven Modeling of Adsorption Energy in Heterogeneous Catalysis, arXiv, preprint, arXiv:2405.20397, 2024, DOI: 10.48550/arXiv.2405.20397.
25 H. Mai, T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, Machine Learning for Electrocatalyst and Photocatalyst Design and Discovery, *Chem. Rev.*, 2022, **122**(16), 13478–13515.
26 S. Pablo-García, S. Morandi, R. A. Vargas-Hernández, K. Jorner, Ž. Ivković, N. López and A. Aspuru-Guzik, Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks, *Nat. Comput. Sci.*, 2023, **3**(5), 433–442.
27 S. Wu, Z. Wang, H. Zhang, J. Cai and J. Li, Deep learning accelerates the discovery of two-dimensional catalysts for hydrogen evolution reaction, *Energy Environ. Mater.*, 2023, **6**(1), e12259.
28 Y. Bian, Y. Wang, Z. Yang, Z. Yin, H. Zhao, Y. Liu, H. Shi, Y. Su, Y. Deng and H. Wang, Accelerating the Prediction of g-C3N4-Supported Dual-Atom Catalysts for Photocatalytic CO2 Reduction to CO and HCOOH: A Machine Learning and DFT Combined Approach, *Adv. Energy Mater.*, 2025, **15**(39), e03855.
29 W. Xu, E. Diesen, T. He, K. Reuter and J. T. Margraf, Discovering high entropy alloy electrocatalysts in vast composition spaces with multiobjective optimization, *J. Am. Chem. Soc.*, 2024, **146**(11), 7698–7707.
30 S. Ma, S. Zheng, W. Zhang, D. Chen and F. Pan, Algebraic graph-based machine learning model for Li-cluster prediction, *J. Phys. Chem. A*, 2023, **127**(8), 2051–2059.
31 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery, *Matter*, 2021, **4**(5), 1578–1597.
32 H. Li, X. Li, P. Wang, Z. Zhang, K. Davey, J. Q. Shi and S.-Z. Qiao, Machine learning big data set analysis reveals C–C electro-coupling mechanism, *J. Am. Chem. Soc.*, 2024, **146**(32), 22850–22858.
33 A. S. Nair, L. Foppa and M. Scheffler, Materials-discovery workflow guided by symbolic regression for identifying acid-stable oxides for electrocatalysis, *npj Comput. Mater.*, 2025, **11**(1), 150.
34 X. Liu and H.-J. Peng, Toward next-generation heterogeneous catalysts: empowering surface reactivity prediction with machine learning, *Engineering*, 2024, **39**, 25–44.
35 C. Bozal-Ginesta, S. Pablo-García, C. Choi, A. Tarancón and A. Aspuru-Guzik, Developing machine learning for heterogeneous catalysis with experimental and computational data, *Nat. Rev. Chem.*, 2025, **9**(9), 601–616.
36 L. G. de Araujo, L. Vilcocq, P. Fongarland and Y. Schuurman, Recent developments in the use of machine learning in catalysis: A broad perspective with applications in kinetics, *Chem. Eng. J.*, 2025, **508**, 160872.
37 S. Choung, W. Park, J. Moon and J. W. Han, Rise of machine learning potentials in heterogeneous catalysis: Developments, applications, and prospects, *Chem. Eng. J.*, 2024, **494**, 152757.
38 J. Yin, W. Li, H. Chen, J. Qiu, H. Feng, X. Xu, Q. Jin and X. Wang, CaTS: Toward Scalable and Efficient Transition State Screening for Catalyst Discovery, *ACS Catal.*, 2025, **15**(18), 15754–15764.
39 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, Less is more: Sampling chemical space with active learning, *J. Chem. Phys.*, 2018, **148**(24), 241733.
40 F. Musil, M. J. Willatt, M. A. Langovoy and M. Ceriotti, Fast and accurate uncertainty estimation in chemical machine learning, *J. Chem. Theory Comput.*, 2019, **15**(2), 906–915.
41 D. Tang, R. Ketkaew and S. Luber, Machine learning interatomic potentials for heterogeneous catalysis, *Chem. – Eur. J.*, 2024, **30**(60), e202401148.
42 B. W. Chen, X. Zhang and J. Zhang, Accelerating explicit solvent models of heterogeneous catalysts with machine learning interatomic potentials, *Chem. Sci.*, 2023, **14**(31), 8338–8354.
43 Z. Lian, F. Dattila and N. López, Stability and lifetime of diffusion-trapped oxygen in oxide-derived copper CO2 reduction electrocatalysts, *Nat. Catal.*, 2024, **7**(4), 401–411.
44 K. Tran and Z. W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution, *Nat. Catal.*, 2018, **1**(9), 696–703.
45 R. Ding, J. Chen, Y. Chen, J. Liu, Y. Bando and X. Wang, Unlocking the potential: machine learning applications in electrocatalyst design for electrochemical hydrogen energy transformation, *Chem. Soc. Rev.*, 2024, **53**, 11390–11461.
46 V. Fung, G. Hu, P. Ganesh and B. G. Sumpter, Machine learned features from density of states for accurate adsorption energy prediction, *Nat. Commun.*, 2021, **12**(1), 88.
47 D. Newns, Self-consistent model of hydrogen chemisorption, *Phys. Rev.*, 1969, **178**(3), 1123.
48 B. Hammer and J. K. Norskov, Why gold is the noblest of all the metals, *Nature*, 1995, **376**(6537), 238–240.
49 B. Hammer, Y. Morikawa and J. K. Nørskov, CO chemisorption at metal surfaces and overlayers, *Phys. Rev. Lett.*, 1996, **76**(12), 2141.
50 A. Ruban, B. Hammer, P. Stoltze, H. L. Skriver and J. K. Nørskov, Surface electronic structure and reactivity of transition and noble metals, *J. Mol. Catal. A: Chem.*, 1997, **115**(3), 421–429.
51 M. Mavrikakis, B. Hammer and J. K. Nørskov, Effect of strain on the reactivity of metal surfaces, *Phys. Rev. Lett.*, 1998, **81**(13), 2819.
52 J. K. Nørskov, T. Bligaard, A. Logadottir, J. Kitchin, J. G. Chen, S. Pandelov and U. Stimming, Trends in the exchange current for hydrogen evolution, *J. Electrochem. Soc.*, 2005, **152**(3), J23.

53 A. Nilsson, L. G. M. Pettersson, B. Hammer, T. Bligaard, C. H. Christensen and J. K. Nørskov, The electronic structure effect in heterogeneous catalysis, *Catal. Lett.*, 2005, **100**(3–4), 111–114.
54 J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, Density functional theory in surface chemistry and catalysis, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(3), 937–943.
55 R. S. Olson; R. J. Urbanowicz; P. C. Andrews; N. A. Lavender; L. C. Kidd and J. H. Moore Automating biomedical data science through tree-based pipeline optimization. In *European conference on the applications of evolutionary computation*, 2016, Springer: pp. 123–137.
56 T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Phys. Rev. Lett.*, 2018, **120**(14), 145301.
57 S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran and Z. W. Ulissi, Convolutional Neural Network of Atomic Surface Structures To Predict Binding Energies for High-Throughput Screening of Catalysts, *J. Phys. Chem. Lett.*, 2019, **10**(15), 4401–4408.
58 S. Yue, D. Li, A. Zhang, Y. Yan, H. Yan, Z. Feng and W. Wang, Rational design of single transition-metal atoms anchored on a PtSe2 monolayer as bifunctional OER/ORR electrocatalysts: a defect chemistry and machine learning study, *J. Mater. Chem. A*, 2024, **12**(9), 5451–5463.
59 J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff and J. K. Nørskov, Computational high-throughput screening of electrocatalytic materials for hydrogen evolution, *Nat. Mater.*, 2006, **5**(11), 909–913.
60 X. Guo, S. Lin, J. Gu, S. Zhang, Z. Chen and S. Huang, Simultaneously achieving high activity and selectivity toward two-electron O2 electroreduction: the power of single-atom catalysts, *ACS Catal.*, 2019, **9**(12), 11042–11054.
61 Y. Yang, X. Zhao, T. Liu, Y. Zhang, Y. Hu, X. Liu, G. Wang, D. Wang, J. Bi and Z. Luo, The rational co-doping strategy of transition metal and non-metal atoms on g-CN for highly efficient hydrogen evolution by DFT and machine learning, *Int. J. Hydrogen Energy*, 2024, **56**, 949–958.
62 L. Zhang, X. Guo, S. Zhang and S. Huang, Building up the ''Genome'' of bi-atom catalysts toward efficient HER/OER/ORR, *J. Mater. Chem. A*, 2022, **10**(21), 11600–11612.
63 L. Xu, Y. Huang, H. Lin, X. Wei and F. Ma, Machine learning driven high-throughput screening of S and N-coordinated SACs for eNRR, *Nano Res.*, 2025, **18**(4), 94907289.
64 V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, Gaussian process regression for materials and molecules, *Chem. Rev.*, 2021, **121**(16), 10073–10141.
65 A. P. Bartók; R. Kondor and G. Csányi, On representing chemical environments, arXiv, preprint, arXiv:1209.3140, 2012, DOI: 10.48550/arXiv.1209.3140.
66 M. O. Jäger, E. V. Morooka, F. Federici Canova, L. Himanen and A. S. Foster, Machine learning hydrogen adsorption on nanoclusters through structural descriptors. npj Computational Materials, *npj Comput. Mater.*, 2018, **4**, 37.
67 M. J. Nielsen; L. H. Kempen; J. D. N. Ravn; R. Cheula and M. Andersen Interpretable machine learned predictions of adsorption energies at the metal–oxide interface, arXiv, preprint, arXiv:2505.21428, DOI: 10.48550/arXiv.2505.21428, 2025.
68 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Phys. Rev. Lett.*, 2010, **104**(13), 136403.
69 S. Y. Willow, A. Hajibabaei, M. Ha, D. C. Yang, C. W. Myung, S. K. Min, G. Lee and K. S. Kim, Sparse Gaussian process based machine learning first principles potentials for materials simulations: Application to batteries, solar cells, catalysts, and macromolecular systems, *Chem. Phys. Rev.*, 2024, **5**(4), 041307.
70 S. Y. Willow, D. G. Kim, R. Sundheep, A. Hajibabaei, K. S. Kim and C. W. Myung, Active sparse Bayesian committee machine potential for isothermal–isobaric molecular dynamics simulations, *Phys. Chem. Chem. Phys.*, 2024, **26**(33), 22073–22082.
71 G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang and L. Yang, Physics-informed machine learning, *Nat. Rev. Phys.*, 2021, **3**(6), 422–440.
72 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, *Phys. Rev. Mater.*, 2018, **2**(8), 083802.
73 R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler and L. M. Ghiringhelli, Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO, *J. Phys. Mater.*, 2019, **2**(2), 024002.
74 J. Wang, H. Xie, Y. Wang and R. Ouyang, Distilling accurate descriptors from multi-source experimental data for discovering highly active perovskite OER catalysts, *J. Am. Chem. Soc.*, 2023, **145**(20), 11457–11465.
75 V. Fung, G. Hu, Z. Wu and D.-E. Jiang, Descriptors for hydrogen evolution on single atom catalysts in nitrogen-doped graphene, *J. Phys. Chem. C*, 2020, **124**(36), 19571–19578.
76 L. Foppa, F. Rüther, M. Geske, G. Koch, F. Girgsdies, P. Kube, S. J. Carey, M. Hävecker, O. Timpe and A. V. Tarasov, Data-centric heterogeneous catalysis: identifying rules and materials genes of alkane selective oxidation, *J. Am. Chem. Soc.*, 2023, **145**(6), 3427–3442.
77 X. Lin, X. Du, S. Wu, S. Zhen, W. Liu, C. Pei, P. Zhang, Z.-J. Zhao and J. Gong, Machine learning-assisted dual-atom sites design with interpretable descriptors unifying electrocatalytic reactions, *Nat. Commun.*, 2024, **15**(1), 8169.
78 X. Jiang; G. Liu; J. Xie and Z. Hu Boosting SISSO performance on small sample datasets by using Random Forests prescreening for complex feature selection. arXiv, preprint, arXiv:2409.19209, DOI: 10.48550/arXiv.2409.19209, 2024.
79 W. Shu, J. Li, J.-X. Liu, C. Zhu, T. Wang, L. Feng, R. Ouyang and W.-X. Li, Structure sensitivity of metal catalysts revealed by interpretable machine learning and

first-principles calculations, *J. Am. Chem. Soc.*, 2024, **146**(12), 8737–8745.
80 S. Xi, P. Zhao, C. He and W. Zhang, High-throughput screening of single-atom catalysts on 1 T-TMD for highly active and selective CO2 reduction reaction: Computational and machine learning insights, *J. Catal.*, 2024, **436**, 115610.
81 G. Yin, H. Zhu, S. Chen, T. Li, C. Wu, S. Jia, J. Shang, Z. Ren, T. Ding and Y. Li, Machine learning-assisted high-throughput screening for electrocatalytic hydrogen evolution reaction, *Molecules*, 2025, **30**(4), 759.
82 Y. Zhang, C. Hu, D. Tang and Z. Su, Machine-learning assisted screening of MXene-supported single-atom catalysts for oxygen reduction, *J. Mater. Chem. A*, 2025, **13**(39), 33897–33906.
83 V. Dufaud and M. E. Davis, Design of heterogeneous catalysts via multiple active site positioning in organic–inorganic hybrid materials, *J. Am. Chem. Soc.*, 2003, **125**(31), 9403–9413.
84 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data*, 2014, **1**(1), 1–7.
85 M. De Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta and A. Gamst, A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds, *Sci. Rep.*, 2016, **6**(1), 34256.
86 S. Sanyal; J. Balachandran; N. Yadati; A. Kumar; P. Rajagopalan; S. Sanyal and P. Talukdar Mt-cgcnn: Integrating crystal graph convolutional neural network with multitask learning for material property prediction, arXiv, preprint, arXiv:1811.05660, 2018, DOI: 10.48550/arXiv.1811.05660.
87 L. Chen, Y. Tian, X. Hu, S. Yao, Z. Lu, S. Chen, X. Zhang and Z. Zhou, A universal machine learning framework for electrocatalyst innovation: a case study of discovering alloys for hydrogen evolution reaction, *Adv. Funct. Mater.*, 2022, **32**(47), 2208418.
88 X. Jia and H. Li, Machine learning enabled exploration of multicomponent metal oxides for catalyzing oxygen reduction in alkaline media, *J. Mater. Chem. A*, 2024, **12**(21), 12487–12500.
89 Q. Yu, N. Ma, C. Leung, H. Liu, Y. Ren and Z. Wei, AI in single-atom catalysts: a review of design and applications, *J. Mater. Inform.*, 2025, **5**(1), 9.
90 H. Chun, J. R. Lunger, J. K. Kang, R. Gómez-Bombarelli and B. Han, Active learning accelerated exploration of single-atom local environments in multimetallic systems for oxygen electrocatalysis, *npj Comput. Mater.*, 2024, **10**(1), 246.
91 M. Kim, B. C. Yeo, Y. Park, H. M. Lee, S. S. Han and D. Kim, Artificial intelligence to accelerate the discovery of N2 electroreduction catalysts, *Chem. Mater.*, 2019, **32**(2), 709–720.
92 J. Zheng, S. Wang, S. Deng, Z. Yao, J. Hu and J. Wang, Accelerating the Screening of Modified MA2Z4 Catalysts for Hydrogen Evolution Reaction by Deep Learning-Based Local Geometric Analysis, *Energy Environ. Mater.*, 2024, **7**(6), e12743.
93 S. Back, K. Tran and Z. W. Ulissi, Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning, *ACS Catal.*, 2019, **9**(9), 7651–7659.
94 C. Ai, S. Han, X. Yang, T. Vegge and H. A. Hansen, Graph neural network-accelerated multitasking genetic algorithm for optimizing Pd x Ti1–x H y surfaces under various CO2 reduction reaction conditions, *ACS Appl. Mater. Interfaces*, 2024, **16**(10), 12563–12572.
95 S. Zheng, H. Ding, X. Yang, S. Li and F. Pan, Automating discovery of electrochemical urea synthesis reaction paths via active learning and graph theory, *CCS Chem.*, 2025, **7**(9), 2822–2834.
96 K. Bang, D. Hong, Y. Park, D. Kim, S. S. Han and H. M. Lee, Machine learning-enabled exploration of the electrochemical stability of real-scale metallic nanoparticles, *Nat. Commun.*, 2023, **14**(1), 3004.
97 G. Xu, Y. Xue, X. Geng, X. Hou and J. Xu, A machine learning-based crystal graph network and its application in development of functional materials, *Mater. Genome Eng. Adv.*, 2024, **2**(3), e38.
98 Q. Zhou, H. Shou, S. Qiao, Y. Cao, P. Zhang, S. Wei, S. Chen, X. Wu and L. Song, Analyzing the active site and predicting the overall activity of alloy catalysts, *J. Am. Chem. Soc.*, 2024, **146**(22), 15167–15175.
99 A. Hashemi, S. Bougueroua, M.-P. Gaigeot and E. A. Pidko, ReNeGate: A reaction network graph-theoretical tool for automated mechanistic studies in computational homogeneous catalysis, *J. Chem. Theory Comput.*, 2022, **18**(12), 7470–7482.
100 K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 992–1002.
101 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, *Chem. Mater.*, 2019, **31**(9), 3564–3572.
102 J. Gasteiger; J. Groß and S. Günnemann Directional message passing for molecular graphs, arXiv, preprint, arXiv: 2003.03123, 2020, DOI: 10.48550/arXiv.2003.03123.
103 K. Choudhary and B. DeCost, Atomistic line graph neural network for improved materials property predictions, *npj Comput. Mater.*, 2021, **7**(1), 185.
104 X. Li, R. Chiong, Z. Hu and A. J. Page, A graph neural network model with local environment pooling for predicting adsorption energies, *Comput. Theor. Chem.*, 2023, **1226**, 114161.
105 T. Kaczynski; K. Mischaikow and M. Mrozek, *Computational homology*, Springer Science & Business Media, 2006.
106 S. Lefschetz Algebraic topology, American Mathematical Soc., 1942.

107 E. H. Spanier, *Algebraic topology*, Springer Science & Business Media, 2012.
108 A. D. Smith, P. Dłotko and V. M. Zavala, Topological data analysis: concepts, computation, and applications in chemical engineering, *Comput. Chem. Eng.*, 2021, **146**, 107202.
109 H.-S. Liu, X.-M. Zhang, G.-H. Liang, S. Zheng and J.-F. Li, Investigation of water structure and proton transfer within confined graphene by ab initio molecule dynamics and multiscale data analysis, *Chin. J. Struct. Chem.*, 2025, **44**(6), 100596.
110 S. Zheng, H. Ding, S. Li, D. Chen and F. Pan, Application of topology-based structure features for machine learning in materials science, *Chin. J. Struct. Chem.*, 2023, **42**(7), 100120.
111 G. Carlsson, Topological methods for data modelling, *Nat. Rev. Phys.*, 2020, **2**(12), 697–708.
112 J. Grbić, J. Wu, K. Xia and G.-W. Wei, Aspects of topological approaches for data science, *Found. Data Sci.*, 2022, **4**(2), 165.
113 Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue and Y. Nishiura, Hierarchical structures of amorphous solids characterized by persistent homology, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**(26), 7035–7040.
114 M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow and V. Nanda, A topological measurement of protein compressibility, *Japan J. Ind. Appl. Math.*, 2015, **32**(1), 1–17.
115 K. Xia and G. W. Wei, Persistent homology analysis of protein structure, flexibility, and folding, *Int. J. Numer. Methods Biomed. Eng.*, 2014, **30**(8), 814–844.
116 P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas and V. S. Pande, Persistent voids: a new structural metric for membrane fusion, *Bioinformatics*, 2007, **23**(14), 1753–1759.
117 Y. Yang, S. Guo, S. Li, Y. Wu and Z. Qiao, Topological Data Analysis Combined with High-Throughput Computational Screening of Hydrophobic Metal–Organic Frameworks: Application to the Adsorptive Separation of C3 Components, *Nanomaterials*, 2024, **14**(3), 298.
118 S. Shekhar and C. Chowdhury, Topological data analysis enhanced prediction of hydrogen storage in metal–organic frameworks (MOFs), *Mater. Adv.*, 2024, **5**(2), 820–830.
119 G.-H. Liang, H.-S. Liu, X.-M. Zhang, J.-F. Li and S. Zheng, Topology-based machine learning for predicting curvature effects in metal-nitrogen-carbon single-atom catalysts, *J. Energy Chem.*, 2025, **105**, 608–616.
120 B. Wang, S. Zheng, J. Wu, J. Li and F. Pan, Inverse design of catalytic active sites via interpretable topology-based deep generative models, *npj Comput. Mater.*, 2025, **11**(1), 147.
121 S. Zheng, X.-M. Zhang, H.-S. Liu, G.-H. Liang, S.-W. Zhang, W. Zhang, B. Wang and J. Yang, Jin, X. a.; Pan, F. Active phase discovery in heterogeneous catalysis via topology-guided sampling and machine learning, *Nat. Commun.*, 2025, **16**(1), 1–13.
122 J. Cui; F. Wu; H. Zhao; M. Feng; X. Evangelopoulos; A. I. Cooper and Y. Choi L 2M 3OF: A Large Language Multimodal Model for Metal-Organic Frameworks, arXiv, preprint, arXiv:2510.20976, DOI: 10.48550/arXiv.2510.20976, 2025.
123 D. H. Mok and S. Back, Generative pretrained transformer for heterogeneous catalysts, *J. Am. Chem. Soc.*, 2024, **146**(49), 33712–33722.
124 Y.-L. Liao; B. Wood; A. Das and T. Smidt Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, arXiv, preprint, arXiv:2306.12059, DOI: 10.48550/arXiv.2306.12059, 2023.
125 G. Yang, S. Jiang, Y. Luo, S. Wang and J. Jiang, Cross-modal prediction of spectral and structural descriptors via a pretrained model enhanced with chemical insights, *J. Phys. Chem. Lett.*, 2024, **15**(34), 8766–8772.
126 L. Zhao, M. Zhou and J. Jiang, Machine Learning-Based Multispectral Fusion for Analyzing Molecular Structural Features, *J. Phys. Chem. Lett.*, 2025, **16**(18), 4382–4391.
127 T. Hu, Z. Zou, B. Li, T. Zhu, S. Gu, J. Jiang, Y. Luo and W. Hu, Deep Learning for Bidirectional Translation between Molecular Structures and Vibrational Spectra, *J. Am. Chem. Soc.*, 2025, **147**(31), 27525–27536.
128 T. Yang, D. Zhou, S. Ye, X. Li, H. Li, Y. Feng, Z. Jiang, L. Yang, K. Ye and Y. Shen, Catalytic structure design by AI generating with spectroscopic descriptors, *J. Am. Chem. Soc.*, 2023, **145**(49), 26817–26823.
129 S. Sun, P. Wang, Y. He, J. Yang and S. Li, Multimodal Cross-Attention Molecular Property Prediction for Text, Sequence, Graph, and Geometry, *ACS Omega*, 2025, **10**(46), 56225–56239.
130 P. Rocabert-Oriols, C. L. Conte, N. López and J. Heras-Domingo, Multi-modal contrastive learning for chemical structure elucidation with VibraCLIP, *Digital Discovery*, 2025, **4**(12), 3818–3827.
131 Y. Luo and L. Deng, MolCL-SP: a multimodal contrastive learning framework with non-overlapping substructure perturbations for molecular property prediction, *Bioinformatics*, 2025, **41**(10), btaf507.
132 C.-Y. Liu and T. P. Senftle, Finding physical insights in catalysis with machine learning, *Curr. Opin. Chem. Eng.*, 2022, **37**, 100832.
133 Z. Wang, X. Liu, H. Chen, T. Yang and Y. He, Exploring multi-fidelity data in materials science: Challenges, applications, and optimized learning strategies, *Appl. Sci.*, 2023, **13**(24), 13176.
134 S. M. Goodlett, J. M. Turney and H. F. Schaefer, Comparison of multifidelity machine learning models for potential energy surfaces, *J. Chem. Phys.*, 2023, **159**(4), 0021–9606.
135 V. Vinod and P. Zaspel, QeMFi: A multifidelity dataset of quantum chemical properties of diverse molecules, *Sci. Data*, 2025, **12**(1), 202.
136 J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang and S. Han, Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials, *J. Am. Chem. Soc.*, 2024, **147**(1), 1042–1054.