Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d6sc01469k

All publication charges for this article have been paid for by the Royal Society of Chemistry

Generative intelligence explores the chemical space of ten million catalysts

Ruili Li,[af] Shuoqi Zhang,[af] Qingli Tang,[bc] Qingqing Mao,[de] Ritankar Das,[e] Rui Qi,*[bc] Beien Zhu*[bc] and Yi Gao*[bc]

Discovery of catalytic materials requires systematic exploration of vast chemical spaces. However, the scope of exploration that can be achieved using conventional theoretical and experimental methods is very limited. Herein, we present a scalable framework based on distributed generative transformers, which integrates a transformer-based generative model with a distributed, parallel generation-screening workflow. By coupling dimensionality reduction with a machine-learning-potential (MLP) model for performance prediction, we construct a catalyst structure library comprising over ten million candidates-expanding the accessible design space by two orders of magnitude relative to existing generative models. Leveraging a pretrained model followed by task-specific fine-tuning, the tailored conditional generation strategy achieves >90% validity for target adsorbates such as CH_3, thereby enabling focused exploration of methane conversion catalysts. Machine learning-accelerated screening of this massive library efficiently identifies 26 known active catalysts and more than 1200 previously unreported candidates. Subsequent subgroup discovery (SGD) analysis reveals synergistic elemental effects that modulate the surface electronic structure, tuning CH_3 binding energies into the optimal window for CH_4 activation. This work establishes a generalizable paradigm that seamlessly bridges generative exploration with high-throughput performance mapping, dramatically accelerating catalyst discovery across unprecedented chemical spaces.

Received 20th February 2026
Accepted 13th May 2026

DOI: 10.1039/d6sc01469k

rsc.li/chemical-science

1 Introduction

The discovery of catalytic materials is pivotal for the sustainable production of chemicals and fuels, yet it typically involves navigating vast chemical spaces encompassing tens to hundreds of millions of possible compositions and structures.[1,2] Traditional catalyst development has largely relied on empirical trial-and-error approaches, which are inherently slow and resource-intensive.[3] High-throughput experimental and computational strategies have partially addressed these limitations by enabling rapid, large-scale screening of catalysts.[4–7] More recently, machine learning (ML) has emerged as a powerful paradigm for catalyst discovery. Compared with conventional high-throughput approaches, ML algorithms—including support vector machines (SVMs),[8] artificial neural networks (ANNs),[9] and decision trees[10]—can uncover hidden correlations and identify key descriptors from high-dimensional experimental and computational datasets, thereby reducing dependence on chemical intuition and accelerating the identification of potential catalyst candidates.[11–20] For example, Deng *et al.* implemented an automated ML workflow that integrates descriptor extraction, model optimization, and large-scale screening, successfully evaluating 10 950 single-atom alloy surfaces and identifying several highly active catalysts for methane cracking.[21]

Recent advances in artificial intelligence, particularly generative models, offer a powerful new avenue for materials design.[22] Generative approaches, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, have demonstrated the ability to produce novel, chemically plausible materials and to enable inverse design of targeted structures such as zeolites, cubic semiconductors, and metal–organic frameworks.[23–32] Transformer-based large language models (LLMs) have further enabled the direct generation of material structures from text.[33–39] Notably, CrystaLLM learns directly from CIF text to generate diverse, structurally valid inorganic crystals, exhibiting strong generalization to unseen compositions and space groups.[40] In the catalytic domain, transformer-based generative models have

[a]Key Laboratory of Interfacial Physics and Technology, Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201800, China
[b]Photon Science Research Center for Carbon Dioxide, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China. E-mail: qir@sari.ac.cn; zhube@sari.ac.cn; gaoyi@sari.ac.cn
[c]State Key Laboratory of Low Carbon Catalysis and Carbon Dioxide Utilization, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China
[d]Incept Labs, Hayward, CA 94541, USA
[e]Titan Holdings, Hayward, CA 94541, USA
[f]University of Chinese Academy of Sciences, Beijing 100049, China

demonstrated significant potential. Mok and Back applied the GPT-2-based architecture to generate heterogeneous catalyst surfaces, bridging structural generation and property prediction.[41] More recently, frameworks such as MAGECS, combining graph neural networks with global optimization strategies, have enabled efficient property-driven exploration for CO_2 reduction catalysts.[42] Despite these advances, existing approaches typically generate only on the order of hundreds of thousands of catalyst structures, which is far too few to comprehensively explore the vast, high-dimensional chemical space of real catalytic systems. Moreover, a fundamental challenge remains in distilling vast numbers of generated candidates into a few truly transformative and practically viable catalysts, such as those for "holy grail" reactions like methane conversion.

To address these challenges, we propose a scalable generation-screening framework based on distributed generative transformers, combining transformer-based generative modeling with a distributed, parallel generation architecture. This framework enables catalyst discovery at the ten-million-structure scale, expanding the accessible generation capacity of existing models by two orders of magnitude. The platform establishes a fully integrated workflow spanning large-scale structure generation, dimensionality reduction, MLP-assisted property evaluation, and efficient screening. In particular, a tailored conditional generation strategy, initialized from a pretrained model and subsequently fine-tuned, enables directed synthesis of specific adsorbates (*e.g.*, CH_3) with over 90% accuracy. Focusing on CH_4 conversion, this workflow successfully rediscovers approximately 26 experimentally established catalysts and identifies more than 1200 previously unreported candidates with potential activity from 10 million generated structures, substantially expanding the accessible catalyst design space. Furthermore, subgroup discovery (SGD) analysis elucidates the role of synergistic elemental effects in tuning the surface electronic structure toward optimal CH_3 binding energies. These results establish a general and scalable framework that integrates large-scale generative exploration with data-driven performance mapping, thereby enabling the systematic discovery of catalytic materials across diverse reaction systems and adsorbates.

2 Results and discussion

An overview of the generative and screening workflow employed in this work is shown in Fig. 1. First, the transformer-based[43] generative model is pretrained on the two million catalyst structures from the Open Catalyst 2020 Structure to Energy and Forces dataset (OC20-S2EF 2M),[44] which provides structures with diverse adsorbates and a broad spectrum of elemental compositions. A similar model architecture has been validated recently for reliable structure generation.[41] In our model, it consists of a 12-layer architecture with 512 hidden dimensions, 8 attention heads, and a maximum sequence length of 1024. Structures were represented as tokenized strings of lattice parameters, atomic species, and three-dimensional coordinates and validated on the OC20-S2EF Val-ID set. Fine-tuning, which retrains a pretrained model on a domain-specific dataset, was applied to the transformer-based generative model to align catalyst compositions and geometries with the fine-tuning dataset while retaining the generative performance of the original model (see supplementary information Methods). CH_4 conversion remains a central challenge in heterogeneous catalysis, with surface-bound CH_3 serving as a key intermediate across diverse reaction pathways.[45] Although CH_3 binding energy is a simplified descriptor and does not fully capture the full complexity of methane conversion, it provides a practical basis for large-scale screening. Because the initial C–H bond cleavage of CH_4 is widely regarded as one of the most difficult and often rate-limiting steps in methane conversion,[21,46,47] descriptors related to early methane activation are chemically motivated for identifying candidate catalysts. Given that explicit transition-state calculations are computationally prohibitive at the ten-million-structure scale explored here, we employed CH_3 as a first-pass activity descriptor and used targeted fine-tuning on CH_3-adsorbed structures to focus generation on chemically relevant configurations. Motivated by the pivotal role of CH_3 intermediates, we developed the distributed generative transformers framework by combining a CH_3-conditioned Transformer-based generative model with GPU-based distributed parallel generation, which maximizes computational throughput, achieves sustained generation rates of millions of structures per day, and enables the assembly of a ten-million-scale catalyst structure database for CH_4 conversion. Second, UMAP-based dimensionality reduction combined with hash-based deduplication was used to compress and organize the ten-million-scale structural database, removing redundancies while preserving the diversity of CH_3 coordination environments. Clustering then identified high-quality regions, from which representative structures with favorable predicted performance were selected for evaluation with machine-learning potential, narrowing the dataset to approximately one million structures. Third, a machine learning model was applied to assess adsorption binding energies across millions of generated structures, enabling rapid exploration of the structural space and establishing material–performance mapping that serves as the basis for subsequent performance screening. Fourth, to effectively evaluate the activity of these catalysts, we use the binding energy of CH_3, typically the key intermediate in CH_4 conversion, as a descriptor. Guided by the Sabatier principle, we established an optimal binding-energy window for high-throughput screening, enabling the systematic identification of candidate catalysts with favorable activity for CH_4 conversion. To sum up, our workflow establishes a closed loop from large-scale structure generation to performance analysis and high-throughput screening, enabling efficient, global exploration of a vast chemical space and identification of potential catalyst candidates.

2.1. Fine-tuning efficiency and structural landscapes of 10 million CH_3-adsorbed catalysts

To enable targeted generation of CH_3 adsorption structures, we evaluated the effect of training dataset size on the performance of the fine-tuning model and measured the corresponding yield

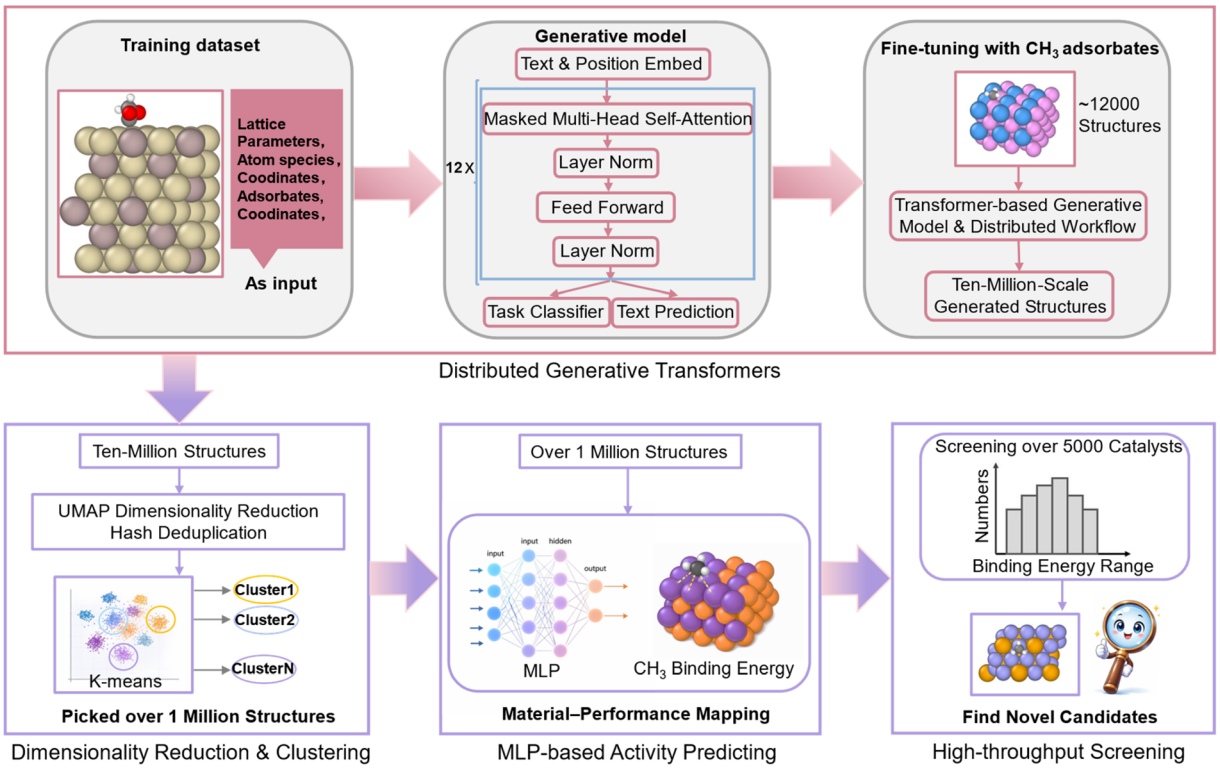

Fig. 1 Schematic overview of the generative AI-driven workflow for large-scale catalyst structure exploration and screening.

of CH_3 adsorption structures. The impact of fine-tuning data size on the generation efficiency of CH_3-adsorbed structures is summarized in Fig. 2a. A pronounced increase in the generation success rate is observed as the number of CH_3-containing structures used for fine-tuning increases from 2000 to 6000, indicating that expanding the high-quality, targeted training dataset enhances the model's ability to learn the complex features of CH_3 adsorption and effectively biases the generative distribution toward the target structures. In this way, the success rate grows and tends to converge. This transition reflects a move from inadequate adsorbate conditioning to a stable generative regime, in which the model reliably incorporates the target motif while preserving pretraining-derived structural priors. Following this trend, all subsequent structure generation and analysis employed the model fine-tuned on 12 000 CH_3-adsorbed structures, at the success-rate plateau, to ensure stable and reliable performance. Leveraging the fine-tuned model and distributed generation framework, we performed large-scale structure generation targeting CH_3 adsorption, yielding 10 000 000 candidate structures, of which 9 999 949 corresponded to valid atomic configurations (as shown in Table 1). Among these, 9 051 967 structures contained CH_3 adsorbates, consistent with the high generation success rate of the fine-tuning model. To ensure structural plausibility, all structures underwent validity checks for overlapping atoms based on atomic distance criteria, with atoms considered overlapping if separated by less than 0.7 Å. This criterion was used as a geometric filter to exclude obviously overlapping unphysical structures before screening. After this filtering step, a total of 8 632 826 non-overlapping CH_3-adsorbed structures were retained for subsequent analysis. This generated dataset expands the catalyst structure space by two orders of magnitude over existing frameworks (see Table S1), offering an unprecedentedly broad basis for exploring adsorption catalysts. To visualize the large-scale generated dataset, one million structures were projected onto a Uniform Manifold Approximation and Projection (UMAP)[48] density map (see supplementary information Methods), with color intensity representing the number of structures per pixel (Fig. 2b). Two dominant high-density clusters are apparent, each containing tens to hundreds of thousands of structures. A continuous bridge connects the clusters, indicating gradual structural transitions, while faint peripheral points reveal rare configurations. Regions 1 and 2 emphasize the internal diversity and high population of the two main clusters.

The elemental composition of the generated structures was further analyzed. We tabulated the element types and their occurrence frequencies in the dataset, presenting the results as a periodic-table heatmap in Fig. 3a. The data show that the generated structures cover a broad range of elements, including both metals and nonmetals, with frequency distributions closely matching the trends observed in the OC20-S2EF 2M dataset (as seen in Fig. S1). At the same time, the generated structures are not simple reproductions of the task-specific fine-tuning data. Structural embedding and compositional analyses (Fig. S2 and S3) reveal clear distributional differences between the generated and fine-tuning datasets, suggesting that the model performs training-informed exploration of catalyst

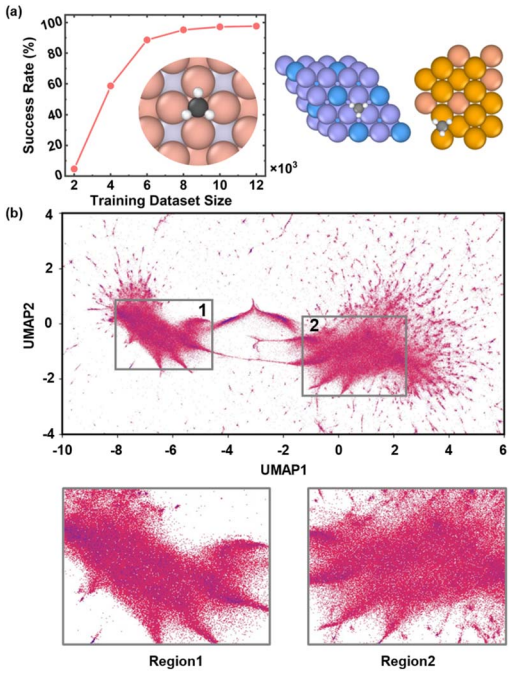

Fig. 2 Fine-tuning efficiency and density mapping of CH_3-adsorbed structures. (a) Success rate of generating CH_3-adsorbed structures as a function of the number of CH_3-containing structures used for the fine-tuning model. Each point represents an independent fine-tuning experiment. The generated structure is shown on the right. (b) Density map of generated structures projected onto a two-dimensional UMAP embedding, where color intensity represents the local structure density. Two dominant high-density regions (region 1 and region 2) are highlighted, corresponding to major structural families in the generated dataset.

chemical space. Consistent with this, SOAP-based analysis of local environments (see supplementary information Methods) shows a broader nearest-neighbor distance distribution for the generated structures, together with substantial populations extending beyond the high-similarity region of the fine-tuning set (Fig. S4 and S5).

To evaluate the overall distribution of the generated structures in feature space, preliminary dimensionality reduction was performed using the UMAP approach. Fig. 3b and S6 show the distribution of randomly selected subsets of generated structures in the two-dimensional UMAP embedding, colored according to the major element fraction and the coordination number within 6 Å of the CH_3 adsorption site, respectively. As illustrated, the generated structures are largely concentrated in a continuous region of the low-dimensional embedding, forming a dominant connected "island". This visualization provides a qualitative view of the continuity and diversity of the generated space in the learned feature representation. The continuous color variation in Fig. 3b indicates gradual compositional changes, reflecting a series of solid solutions or alloys. Similarly, the smooth color gradient in Fig. S6 visualizes differences in atomic coordination environments, showing that the generated structures are locally valid and diverse. Representative structures sampled from different embedding regions are provided in Fig. S7 to illustrate this structural diversity. These results demonstrate that the targeted fine-tuning-based large-scale generative framework enables efficient generation of candidate adsorption catalysts combining chemical diversity and structural validity, while their physical plausibility is evaluated through separate statistical and first-principles analyses.

2.2. Hierarchical screening for million-scale CH_3-adsorbed candidate structures

Building on the large-scale CH_3-adsorbed structural space generated above, hierarchical filtering was applied to over eight million structures using hash-based duplication and clustering to enable subsequent catalytic performance evaluation. Fig. 4a summarizes the hierarchical filtering and screening workflow used to select structures for MLP-based performance evaluation. After atomic-overlap filtering, approximately 8.6 million CH_3-adsorbed structures are retained. Next, a hash-based deduplication scheme (see supplementary information Methods) was employed to remove geometrically redundant configurations with equivalent local adsorption environments. This step preserves the diversity of local coordination motifs while reducing the dataset to about 7.6 million structures. The remaining structures were then grouped by *k*-means clustering based on the local environment descriptors (described in the supplementary information Methods), yielding 2000 clusters. This clustering provides a coarse-grained organization of the generated structural space, facilitating collective analysis of structures with similar adsorption environments. To visualize the organization induced by clustering, a two-dimensional UMAP embedding of a reduced feature space is shown in Fig. 4b. Although clustering was performed in a 64-dimensional Singular Value Decomposition (SVD)-reduced space (see supplementary information Methods),[49] the UMAP projection provides a qualitative visualization of the global structure distribution. Several dense regions, or "islands", are observed, within which multiple *k*-means clusters are interwoven, indicating the presence of closely related local coordination environments within broader regions of structural similarity. To enable efficient MLP-based screening, an initial subset

Table 1 Summary of large-scale structure generation

	Attempted generation	Successfully generated	CH_3-containing	Without overlap
Number of structures	10 000 000	9 999 949	9 051 967	8 632 826
Fraction	—	99.99%	90.52%	86.33%

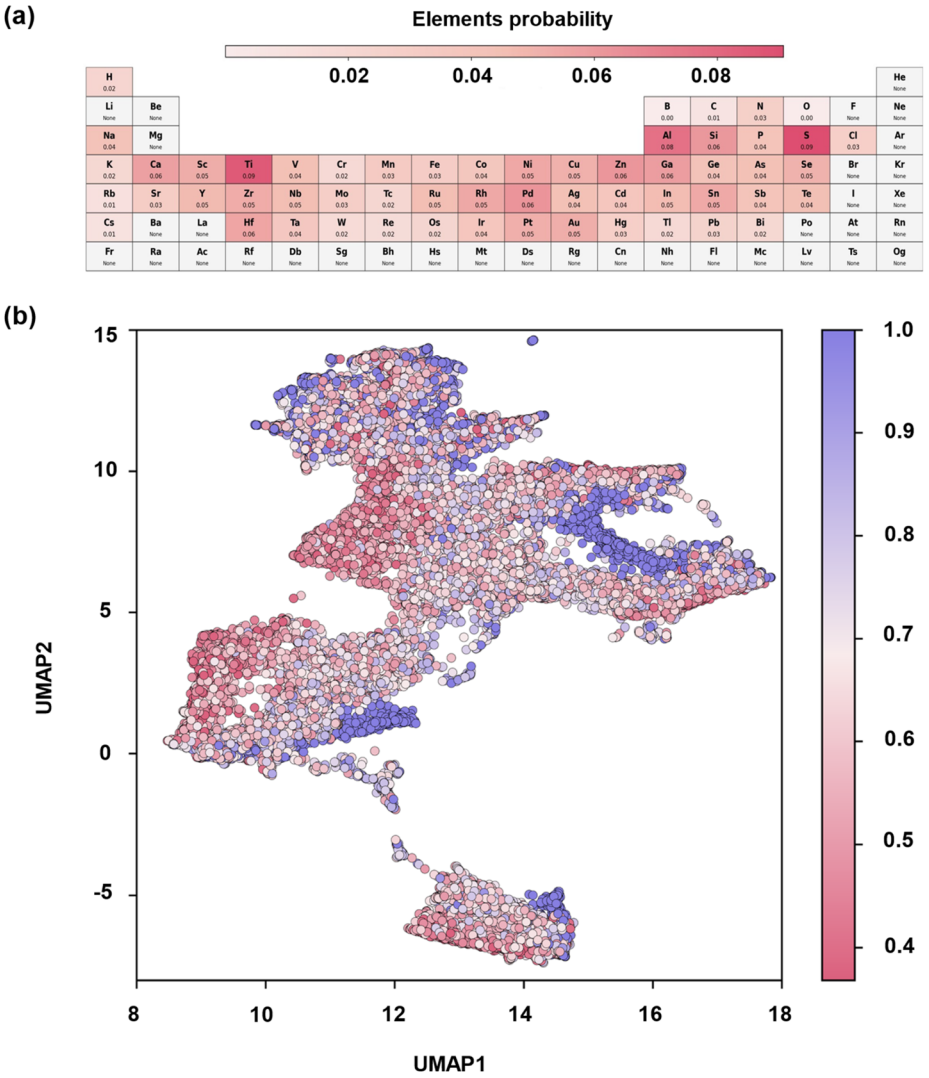

Fig. 3 Elemental diversity and structural distribution of generated CH_3-adsorbed catalysts. (a) Elemental periodic table heatmap showing the occurrence probabilities of elements in the generated structures without atomic overlaps. (b) Two-dimensional UMAP embedding of a randomly sampled subset of generated structures based on local structural descriptors colored by the major element fraction.

corresponding to 1% of the structures from each cluster was randomly sampled and evaluated using the machine learning potential, specifically the pretrained EquiformerV2 model.[50] Clusters were then ranked according to the fraction of structures with negative MLP-predicted binding energies, and the top 400 clusters were selected. All structures contained within these selected clusters, amounting to approximately 1.2 million configurations, were subsequently subjected to full MLP-based evaluation of binding energy of adsorbed CH_3 and used for further catalytic performance analysis. This choice was made in view of computational cost, compressing the generated space to a tractable million-scale subset for subsequent screening while maintaining coverage of structurally and energetically reasonable regions. Here, the MLP model serves as an efficient large-scale screening tool for adsorption energetics rather than a replacement for first-principles calculations. The histogram in Fig. 4c shows the distribution of MLP-predicted binding energies of CH_3 for the sampled structures. Across the dataset, most CH_3 binding energies are concentrated in an intermediate range, with fewer structures at the high- and low-energy extremes. Specifically, structures with binding energies between -1.5 and -0.5 eV are the most abundant, exceeding 500 000. Moving toward higher or lower energies, the number of structures gradually decreases, with the lowest count observed below -3.5 eV, comprising only about 1300 structures.

Since intermediate adsorption states are typically more relevant to catalytic activity, CH_3 species with binding energies above 0.5 eV, essentially unbound, are expected to contribute minimally to the reaction. Therefore, this high-energy region was excluded from the subsequent elemental analysis. Fig. 4d shows the distribution of elemental compositions across different binding-energy intervals. For each interval, the radial bar charts report both the co-occurrence frequency of specific elements with others and the overall frequency of each element

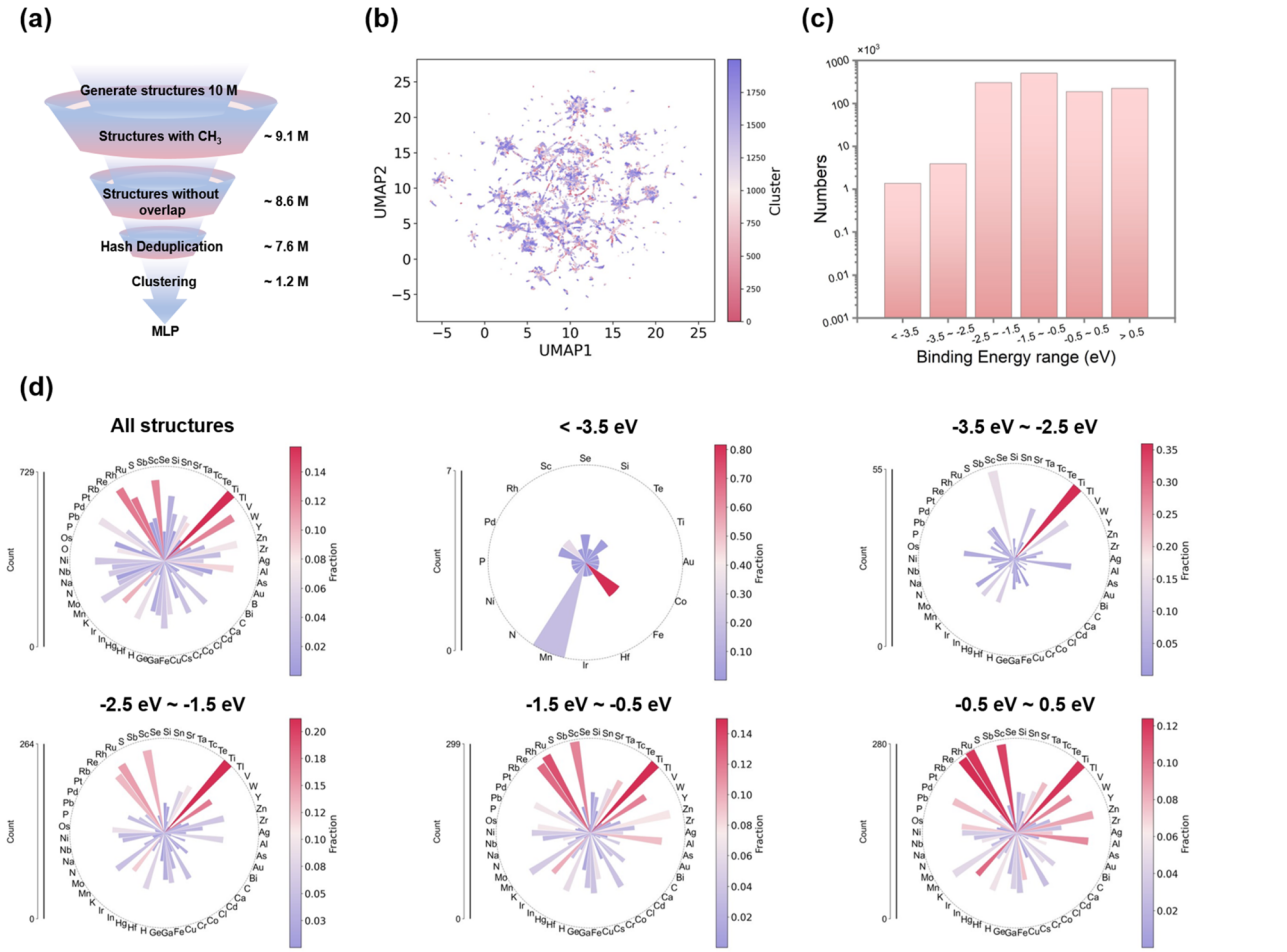

Fig. 4 Large-scale screening and performance evaluation of CH_3-adsorbed structures *via* integrated methods. (a) Schematic illustration of the hierarchical filtering and screening workflow applied to the generated structures. The funnel diagram outlines successive steps from overlap removal and hash-based deduplication to clustering and MLP-based binding-energy evaluation, with numbers indicating the remaining structures at each stage. (b) UMAP embedding of the deduplicated structures clustered using the *k*-means algorithm. Each point represents one structure and is colored according to its assigned cluster ID among the 2000 clusters. (c) Distribution of MLP-predicted energies for the screened structures, shown as a histogram over predefined energy intervals. The vertical axis reports the number of structures within each energy range. (d) Radial bar charts showing element-resolved distributions for structures within MLP-predicted different energy intervals. In each subplot, the radial bar length represents the number of occurrences of this element in combination with other elements, while the color scale encodes the total occurrence frequency of that element across the corresponding subset.

within that energy range. Clear differences in elemental distributions are observed between energy ranges, indicating that the predicted adsorption performance is not uniformly distributed across chemical space but exhibits clear element-dependent trends. For example, Ti is consistently the most frequent element across sub-intervals between −3.5 and −0.5 eV, closely reflecting its overall prevalence in the OC20-S2EF 2M dataset and suggesting that the model preserves compositional features learned during pretraining. In the −1.5 to −0.5 eV interval, Rh and Ru[51–53] preferentially appear in multicomponent combinations, consistent with experimental observations that these elements often catalyze CH_4 conversion in alloyed forms.

Fe-containing structures show the highest fraction in the lowest-energy region (<−3.5 eV). These results demonstrate that the element-energy associations revealed by interval-based analysis both reflect the compositional statistics of the pretraining dataset and correspond to known roles of specific elements in experimental catalytic systems. The integration of MLP-predicted binding energies with large-scale generative screening can effectively identify potentially interesting catalysts towards targeted reactions, providing interpretable compositional guidance for generative catalyst design.

2.3. Composition-driven energy assessment and screening of CH_4 conversion catalysts

To enable statistical analysis and high-throughput screening for CH_4 conversion, the generated structures at the million scale were classified by elemental composition into unary, binary, ternary, and higher-order catalysts, comprising over 5000 distinct material types after excluding combinations containing highly toxic or extremely reactive elements. For each material, the mean CH_3 binding energy and its standard deviations were

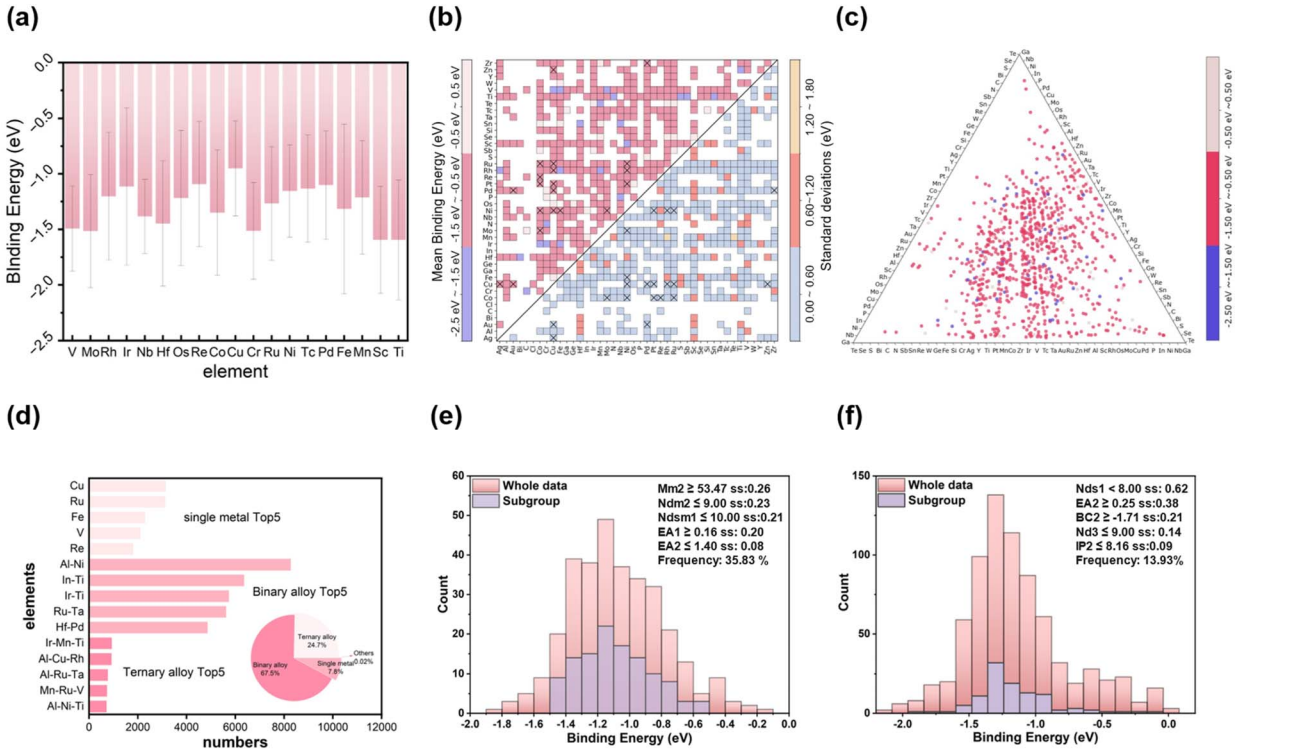

Fig. 5 CH_3 binding-energy-window-based high-throughput screening of catalysts: binding-energy distributions, dominant compositions, and key features. (a) Mean CH_3 binding energies of generated single-component catalysts within the energy range of −5 to 0 eV. Bars represent the average binding energy calculated over all structures corresponding to each single-component catalyst, and error bars indicate the standard deviations within the corresponding structural ensemble. (b) Energy distribution of generated binary catalysts in the same energy range as (a). Upper-left triangular entries indicate mean CH_3 binding energies, while lower-right triangular regions show the standard deviations of the averaged binding energies for each binary composition. (c) Distribution of generated ternary catalysts and the corresponding average binding energies of CH_3. (d) Composition statistics of generated catalysts with binding energies between −1.5 and −0.5 eV. The pie chart shows the relative fractions of unary, binary, and ternary materials, while the bar charts highlight the top five most frequently occurring metal materials within each category. SGD-based feature analysis for (e) binary alloy catalysts and (f) ternary alloy catalysts within the selected energy window, illustrating the relative importance of element features for CH_3 binding energy identified from the subgroup.

computed to evaluate performance and structural diversity. Fig. 5a presents the CH_3 binding energies of single-component materials generated within the −5 to 0 eV range. The mean binding energies are relatively concentrated, spanning ∼−1.59 eV for Ti to ∼−0.95 eV for Cu, highlighting the element-specific effects on CH_3 adsorption. Standard deviations reveal more pronounced differences, ranging from ∼0.33 eV for Nb to ∼0.76 eV for Fe, reflecting the diversity of adsorption geometries generated for each element. A set of experimentally well-established single-component metal catalysts (Ir, Rh, Ru, Ni, and Pd) also lie within this optimal binding-energy window.[54–59] The analysis was extended to binary materials (Fig. 5b), where the upper-left triangle shows mean CH_3 binding energies and the lower-right triangle shows the corresponding standard deviations. This allows simultaneous assessment of the energy tendency and the stability of the average value, thereby identifying compositions that are both stable and catalytically active. Most binary catalysts exhibit moderate mean energies with low standard deviations, indicating stable adsorption, while some combinations show broader distributions. According to the Sabatier principle, the binding strength of the key methane intermediate CH_3 should be moderate; thus, we define −1.5 to −0.5 eV as the optimal binding-energy window. Literature-reported high-activity binary catalysts for CH_4 conversion[51–53,60–75] (pink region) fall within this window and correspond to low-variance regions (blue region), marked by " × " in the figure. To further examine whether this recovery is meaningfully associated with the screening criterion, we analyzed the distribution of recovered experimental catalysts across different CH_3-binding-energy ranges (Table S2). Notably, 26 recovered catalysts fall within the target window of −1.5 to −0.5 eV, whereas only 2 appear below −1.5 eV and 2 above −0.5 eV, indicating a clear enrichment of known catalysts in the targeted screening region. Beyond composition-level recovery, representative generated structures further exhibit literature-consistent local/surface motifs (Fig. S8),[76–78] supporting that the rediscovery is chemically meaningful rather than composition-only. These observations validate the reliability and efficiency of our screening approach, indicating that binary materials within the optimal window are highly likely to exhibit catalytic activity for methane conversion. Fig. 5c shows the CH_3 binding-energy distribution of ternary materials, with each point representing a distinct composition colored by its average binding energy. The ternary map highlights the high diversity of

generated materials, demonstrating that the framework effectively explores the ternary compositional space. Most compositions lie within the intermediate energy range, overlapping with the optimal binding-energy window, and this region also includes experimentally reported ternary systems such as Ni–Fe–Cr, further supporting the relevance of the generated candidates.[79] Thus, these results indicate that the generative framework efficiently produces multi-component catalysts with continuous and physically reasonable CH_3 adsorption properties, providing a rich library for targeted design and high-throughput screening of CH_4 conversion catalysts.

We further analyzed the elemental composition of catalysts within the optimal CH_3 binding-energy window (−1.5 to −0.5 eV, Fig. 5d). Binary materials dominate (∼67%), followed by ternary (∼25%) and unary (∼8%) compositions, with higher-order structures negligible, indicating that the preferred-energy structures are primarily concentrated in binary and ternary compositions. Bar charts highlight the five most frequent compositions in each category after excluding the radioactive element Tc. Unary materials are dominated by common transition metals (Cu, Ru, Fe, V, and Re) that appear most frequently, consistent with their prevalence in experimental catalysis. For binary materials, combinations such as Al–Ni, In–Ti, Ir–Ti, Ru–Ta, and Hf–Pd are particularly abundant, each comprising several thousand structures. Multiple ternary compositions also exceed 800 structures. Overall, these statistics highlight that binary materials form the main component of generated structures within the optimal energy window, while ternary materials further enrich compositional diversity, expanding the design space for catalysts. To identify key descriptors governing CH_3 binding on alloy surfaces, subgroup discovery (SGD) was applied within the optimal binding-energy window, using binary alloys as a representative (Table S3 and Fig. S9). By maximizing the sample within the optimal CH_3 binding-energy window (−1.5 to −0.5 eV), a representative subgroup of 115 points (35.8% of the whole data) was identified, defined by the inequalities M2 ≥ 53.47, Nd2 ≤ 9.00, Nds1 ≤ 10.00, EA1 ≥ 0.16, and EA2 ≤ 1.40, where M, Nd, Nds, and EA denote relative atomic mass, number of d valence electrons, total number of d and s valence electrons, and electron affinity, respectively. Two metal elements in the binary alloy were marked as element1 and element2. Optimal CH_3 binding occurs when a heavy transition metal (*e.g.*, Fe, Co, and Ni) with partially filled d orbitals (element2) serves as the active site, paired with a low-valence or early transition metal (*e.g.*, Al) that modulates the surface electronic structure. Using the SGD screening criteria, known effective CH_4 conversion catalysts—such as Co–Mo, Co–Ru, Co–Ni, Ni–Ru, and Ni–Rh—were successfully recovered.[51,53,62,63,75] Notably, Al–Ni, the most frequently occurring binary alloy with moderate CH_3 binding, also satisfies these criteria and has been experimentally synthesized, further supporting the robustness and practical relevance of the proposed approach.[80] Extending SGD to ternary alloys (Fig. 5f) uncovers a cooperative design motif whereby one electron-modulating element (*e.g.*, Ni, Ru or Co) governs surface electronic descriptors (EA, BC, and IP), whereas two bandwidth-modulating elements (*e.g.*, Al/Ta, Al/Ti or Mn/V) collectively modulate d-band filling and width through Nd and Nds. On this basis, ternary combinations such as Al–Ru–Ta, Al–Ni–Ti, and Mo–Ru–V emerge as potential candidates for CH_3 binding optimization. Fig. S10 shows representative structures of unary, binary, and ternary catalysts within the optimal binding-energy window.

Overall, following the above screening workflow and excluding known catalysts and compositions containing toxic or radioactive elements, we identified over 1200 previously unreported candidates with potential catalytic activity for CH_4 conversion, offering new directions for experimental validation and highlighting the power of our generative and screening framework to uncover potential CH_4-conversion candidate systems. The present framework should therefore be viewed as a high-throughput candidate-generation and prioritization

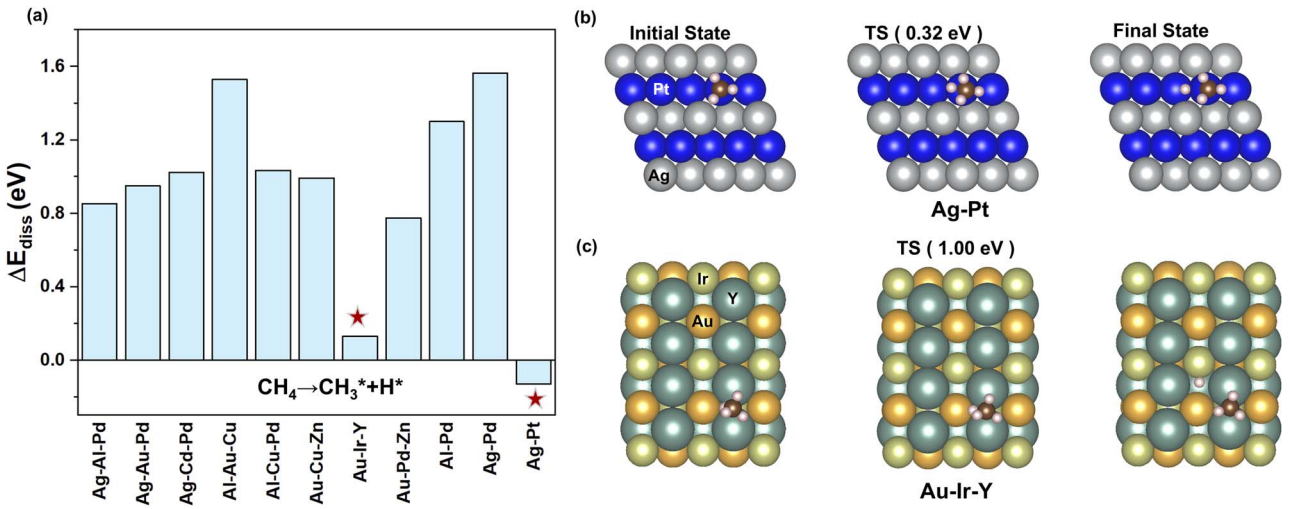

Fig. 6 (a) The reaction energies (ΔE_{diss}) for the first CH_4 dissociation step on representative screened candidates. The five-pointed star represents the system that has been selected for the energy barrier calculation. The optimized initial state, transition state (TS), and final state for $CH_4 \rightarrow CH_3^* + H^*$ on (b) Ag–Pt and (c) Au–Ir–Y surfaces.

strategy, rather than a fully unbiased inverse-design scheme. To further examine the plausibility of representative screened candidates, we surveyed available experimental synthesis precedents and E_{hull} data of the Materials Project for some selected alloy systems (Tables S4 and S5), providing evidence that these potential compositions are connected to experimentally and thermodynamically accessible alloy spaces. Representative DFT calculations further show that selected adsorption structures retain their key adsorption motifs after structural relaxation (Fig. S11), supporting the physical plausibility of the screened candidates. We further assessed the reaction-level relevance of the MLP screening by performing DFT calculations for a separate set of candidates whose predicted CH_3 adsorption energies fall within the target screening window; the corresponding systems and energies are provided in Table S6. As shown in Fig. 6a, the first CH_4 dissociation energies ($CH_4 \rightarrow CH_3^* + H^*$) differ substantially across these systems. The optimized initial state and final state of representative screened candidates are shown in Fig. S12. Based on their relatively favorable energetics, Ag–Pt and Au–Ir–Y were further selected for the energy barrier calculations, and the optimized initial, transition, and final states are shown in Fig. 6b and c. In particular, Ag–Pt exhibits a notably low first-step C–H activation barrier, indicating its promise for CH_4 activation toward CH_3.

From ten million generated candidates, a substantial number of structures exhibiting energetically reasonable CH_3 adsorption have been identified and organized by composition and structural features. These structures constitute a practical candidate pool that can serve as a starting point for further validation. To facilitate reuse and community access, all generated structures together with the associated composition labels and predicted binding energetics have been made publicly available *via* an open repository (see Data availability).

3 Conclusions

In summary, we have developed a scalable generation–screening framework based on distributed generative transformers, which integrates a Transformer-based generative model with distributed parallel computing, dimensionality-reduction analysis, and MLP-based property prediction. This platform enables catalyst discovery at the ten-million-structure scale, representing an increase of two orders of magnitude over existing generative models. A tailored conditional generation strategy, built upon a pretrained model and subsequently fine-tuned, achieves over 90% yield for specific adsorbates, enabling the targeted exploration of a CH_3-adsorbed catalyst structure library for CH_4 conversion. The distributed generative transformers framework, combined with dimensionality reduction and MLP-assisted property evaluation, enables rapid screening of millions of generated structures. Our screening is designed to identify catalysts favorable for the difficult and kinetically important initial C–H activation of CH_4. In this way, the framework successfully reproduces 26 known catalysts and identifies over 1200 previously unreported candidate materials with potential activity for the first C–H bond activation in CH_4 conversion. SGD analysis further demonstrates that synergistic elemental effects tune the surface electronic structure, placing CH_3 binding within the optimal range for CH_4 conversion. All generated structures have been made publicly available, offering an open resource for the catalysis community. Overall, this work establishes a general and extensible platform that bridges high-throughput generative exploration with data-driven performance mapping, enabling systematic, large-scale discovery of catalytic materials from an unprecedentedly large chemical space across diverse reaction systems and adsorbates.

Author contributions

RQ initiated the project. BZ and YG supervised the project. RL performed the calculations and data analysis. RQ wrote the original version. BZ and YG revised the manuscript. All authors participated in the discussions.

Conflicts of interest

There are no conflicts to declare.

Data availability

The codes supporting the findings of this study are publicly available at https://github.com/mosp-catalysis/GenStrucs_CH3. The generated datasets are available at https://doi.org/10.57760/sciencedb.36684. Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6sc01469k.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (92477105 and 92577120), Shanghai Municipal Science and Technology Major Project, and Foundation of the Key Laboratory of Low-Carbon Conversion Science & Engineering, Shanghai Advanced Research Institute, Chinese Academy of Sciences (KLLCCSE-202201Z, SARI, CAS). R. Q. thanks the Innovation Program of Shanghai Advanced Research Institute, CAS (2025CP007). All calculations were performed at the National Supercomputing Center in Shanghai.

Notes and references

1 A. R. Oganov, A. O. Lyakhov and M. Valle, *Acc. Chem. Res.*, 2011, **44**, 227–237.
2 A. Ramirez, E. Lam, D. P. Gutierrez, Y. Hou, H. Tribukait, L. M. Roch, C. Copéret and P. Laveille, *Chem Catal.*, 2024, **4**, 100888.
3 X. Liu, J. Liang, Z. Wang, Q. Li, Y. Deng and H. Wang, *Adv. Energy Mater.*, 2026, **16**, e05497.
4 O. Levy, G. L. W. Hart and S. Curtarolo, *J. Am. Chem. Soc.*, 2010, **132**, 4830–4833.
5 J. Greeley, T. F. Jaramillo, J. Bonde, I. B. Chorkendorff and J. K. Norskov, *Nat. Mater.*, 2006, **5**, 909–913.
6 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191–201.

7 Y. Y. Pan, X. Y. Shan, F. R. Cai, H. Gao, J. N. Xu and M. Zhou, *Angew. Chem., Int. Ed.*, 2024, **63**, e202407116.
8 O. Ivanciuc, *Rev. Comput. Chem.*, 2007, **23**, 291.
9 F. M. Cavalcanti, C. E. Kozonoe, K. A. Pacheco and R. M. de Brito Alves, Application of artificial neural networks to chemical and process engineering, in *Deep Learning Applications*, IntechOpen, 2021.
10 A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, *J. Chemom.*, 2004, **18**, 275–285.
11 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
12 D. T. Ahneman, *Science*, 2018, **360**, 613.
13 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–347.
14 B. Huang and O. A. von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
15 Z. Y. Han, R. H. Gao, T. S. Wang, S. Y. Tao, Y. Y. Jia, Z. J. Lao, M. T. Zhang, J. Q. Zhou, C. Li, Z. H. Piao, X. Zhang and G. M. Zhou, *Nat. Catal.*, 2023, **6**, 1073–1086.
16 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, *Nat. Catal.*, 2022, **5**, 175–184.
17 Z. L. Song, X. Wang, F. T. Liu, Q. H. Zhou, W. J. Yin, H. Wu, W. Q. Deng and J. L. Wang, *Mater. Horiz.*, 2023, **10**, 1651–1660.
18 K. Tran and Z. W. Ulissi, *Nat. Catal.*, 2018, **1**, 696–703.
19 B. C. Weng, Z. L. Song, R. L. Zhu, Q. Y. Yan, Q. D. Sun, C. G. Grice, Y. F. Yan and W. J. Yin, *Nat. Commun.*, 2020, **11**, 3515.
20 M. Zhong, K. Tran, Y. M. Min, C. H. Wang, Z. Y. Wang, C. T. Dinh, P. De Luna, Z. Q. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C. S. Tan, M. Askerka, F. L. Che, M. Liu, A. Seifitokaldani, Y. J. Pang, S. C. Lo, A. Ip, Z. Ulissi and E. H. Sargent, *Nature*, 2020, **581**, 178–183.
21 J. K. Sun, R. Tu, Y. C. Xu, H. Y. Yang, T. Yu, D. Zhai, X. Q. Ci and W. Q. Deng, *Nat. Commun.*, 2024, **15**, 6036.
22 P.-P. De Breuck, H.-C. Wang, G.-M. Rignanese, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2025, **11**, 370.
23 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial nets, *Adv. Neural Inf. Process. Syst.*, 2014, **27**, 2672–2680.
24 T. Xie, X. Fu, O.-E. Ganea, R. Barzilay and T. Jaakkola, *arXiv*, 2021, preprint, arXiv:2110.06197, DOI: 10.48550/arXiv.2110.06197.
25 D. Yan, A. D. Smith and C. C. Chen, *Nat. Comput. Sci.*, 2023, **3**, 572–574.
26 M. Alverson, S. G. Baird, R. Murdock, J. Johnson and T. D. Sparks, *Digital Discovery*, 2024, **3**, 62–80.
27 L. Chen, W. Zhang, Z. Nie, S. Li and F. Pan, *J. Mater. Inf.*, 2021, **1**, 4.
28 B. Kim, S. Lee and J. Kim, *Sci. Adv.*, 2020, **6**, eaax9324.
29 Y. Zhao, M. Al-Fahdi, M. Hu, E. M. D. Siriwardane, Y. Q. Song, A. Nasiri and J. J. Hu, *Adv. Sci.*, 2021, **8**, 2100566.
30 E. M. D. Siriwardane, Y. Zhao, I. Perera and J. J. Hu, *npj Comput. Mater.*, 2022, **8**, 164.
31 Z. P. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr and A. Aspuru-Guzik, *Nat. Mach. Intell.*, 2021, **3**, 76–86.
32 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, Z. L. Wang, A. Shysheya, J. Crabbe, S. Ueda, R. Sordillo, L. X. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C. W. Huang, Z. H. Lu, Y. C. Zhou, H. Yang, H. X. Hao, J. L. Li, C. L. Yang, W. J. Li, R. Tomioka and T. Xie, *Nature*, 2025, **639**, 624–632.
33 A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, *Improving language understanding by generative pre-training*, OpenAI, 2018, https://openai.com/index/language-unsupervised/?utm_source=chatgpt.com.
34 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, **6**, 525–535.
35 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, *Nat. Mach. Intell.*, 2024, **6**, 122–123.
36 T. Xie, Y. Wan, W. Huang, Y. Zhou, Y. Liu, Q. Linghu, S. Wang, C. Kit, C. Grazian and W. Zhang, *arXiv*, 2023, preprint, arXiv:2304.02213, DOI: DOI: 10.48550/arXiv.2304.02213.
37 N. H. Fu, L. Wei, Y. Q. Song, Q. Y. Li, R. Xin, S. S. Omee, R. Z. Dong, E. M. D. Siriwardane and J. J. Hu, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 015001.
38 D. A. Boiko, R. Macknight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
39 D. Flam-Shepherd and A. Aspuru-Guzik, *arXiv*, 2023, preprint, arXiv:2305.05708, DOI: DOI: 10.48550/arXiv.2305.05708.
40 L. M. Antunes, K. T. Butler and R. Grau-Crespo, *Nat. Commun.*, 2024, **15**, 10570.
41 D. H. Mok and S. Back, *J. Am. Chem. Soc.*, 2024, **146**, 33712–33722.
42 Z. L. Song, L. F. Fan, S. H. Lu, C. Y. Ling, Q. H. Zhou and J. L. Wang, *Nat. Commun.*, 2025, **16**, 1053.
43 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, *Language models are unsupervised multitask learners*, OpenAI blog, 2019.
44 L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho and W. Hu, *ACS Catal.*, 2021, **11**, 6059–6072.
45 L. Yan, L. Jiang, C. Qian and S. Zhou, *Energy Rev.*, 2024, **3**, 100065.
46 W. Huang, A. C. Johnston-Peck, T. Wolter, W.-C. D. Yang, L. Xu, J. Oh, B. A. Reeves, C. Zhou, M. E. Holtz, A. A. Herzing, A. M. Lindenberg, M. Mavrikakis and M. Cargnello, *Science*, 2021, **373**, 1518–1523.
47 Y. Wang, P. Hu, J. Yang, Y.-A. Zhu and D. Chen, *Chem. Soc. Rev.*, 2021, **50**, 4299–4358.
48 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018, preprint, arXiv:1802.03426, DOI: DOI: 10.48550/arXiv.1802.03426.
49 G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU press, 2013.
50 Y.-L. Liao, B. Wood, A. Das and T. Smidt, *arXiv*, 2023, preprint, arXiv:2306.12059, DOI: DOI: 10.48550/arXiv.2306.12059.

51 Q. Yin, T. Shen, J. Li, C. Ning, Y. Xue, G. Chen, M. Xu, F. Wang, Y.-F. Song and Y. Zhao, *Chem. Eng. J.*, 2023, **470**, 144416.
52 L. Zhou, J. M. P. Martirez, J. Finzel, C. Zhang, D. F. Swearer, S. Tian, H. Robatjazi, M. Lou, L. Dong and L. Henderson, *Nat. Energy*, 2020, **5**, 61–70.
53 Z. Wu, B. Yang, S. Miao, W. Liu, J. Xie, S. Lee, M. J. Pellin, D. Xiao, D. Su and D. Ma, *ACS Catal.*, 2019, **9**, 2693–2700.
54 M. Danielis, S. Colussi, C. De Leitenburg, L. Soler, J. Llorca and A. Trovarelli, *Angew. Chem.*, 2018, **130**, 10369–10373.
55 W. Hua, Y. C. Dai and H. T. Jiang, *Adv. Mater. Res.*, 2013, **648**, 83–87.
56 T. Kanamori, M. Matsuda and M. Miyake, *Appl. Catal., A*, 2006, **310**, 91–96.
57 Z. Liu, F. Zhang, N. Rui, X. Li, L. Lin, L. E. Betancourt, D. Su, W. Xu, J. Cen and K. Attenkofer, *ACS Catal.*, 2019, **9**, 3349–3359.
58 H. Wang, G. Cui, H. Lu, Z. Li, L. Wang, H. Meng, J. Li, H. Yan, Y. Yang and M. Wei, *Nat. Commun.*, 2024, **15**, 3765.
59 Y. Yao, B. Li, X. Gao, Y. Yang, J. Yu, J. Lei, Q. Li, X. Meng, L. Chen and D. Xu, *Adv. Mater.*, 2023, **35**, 2303654.
60 N. Köpfle, T. Götsch, M. Grünbacher, E. A. Carbonio, M. Hävecker, A. Knop-Gericke, L. Schlicker, A. Doran, D. Kober and A. Gurlo, *Angew. Chem., Int. Ed.*, 2018, **57**, 14613–14618.
61 J. Niu, Y. Wang, S. E. Liland, S. K. Regli, J. Yang, K. R. Rout, J. Luo, M. Rønning, J. Ran and D. Chen, *ACS Catal.*, 2021, **11**, 2398–2411.
62 T. Huang, W. Huang, J. Huang and P. Ji, *Fuel Process. Technol.*, 2011, **92**, 1868–1875.
63 M. Khavarian, S.-P. Chai and A. R. Mohamed, *Fuel*, 2015, **158**, 129–138.
64 C. Palmer, D. C. Upham, S. Smart, M. J. Gordon, H. Metiu and E. W. McFarland, *Nat. Catal.*, 2020, **3**, 83–89.
65 S. M. Kim, P. M. Abdala, T. Margossian, D. Hosseini, L. Foppa, A. Armutlulu, W. van Beek, A. Comas-Vives, C. Copéret and C. Müller, *J. Am. Chem. Soc.*, 2017, **139**, 1937–1949.
66 C. Hammond, M. M. Forde, M. H. Ab Rahim, A. Thetford, Q. He, R. L. Jenkins, N. Dimitratos, J. A. Lopez-Sanchez, N. F. Dummer and D. M. Murphy, *Angew. Chem., Int. Ed.*, 2012, **51**, 5129.
67 N. V. Mdlovu, K.-S. Lin, W.-T. Hong, M. François, A. Hussain and J. Hussain, *J. Energy Inst.*, 2025, 102203.
68 A. Oda, K. Ichino, Y. Yamamoto, T. Ohtsu, W. Shi, Y. Sawada, J. Kumagai, K. Sawabe and A. Satsuma, *J. Am. Chem. Soc.*, 2025, **147**, 30009–30021.
69 Z.-Y. Zhang, T. Zhang, R.-K. Wang, B. Yu, Z.-Y. Tang, H.-Y. Zheng, D. He, T. Xie and Z. Hu, *J. Catal.*, 2022, **413**, 829–842.
70 Z. Jin, L. Wang, E. Zuidema, K. Mondal, M. Zhang, J. Zhang, C. Wang, X. Meng, H. Yang and C. Mesters, *Science*, 2020, **367**, 193–197.
71 N. Agarwal, S. J. Freakley, R. U. McVicker, S. M. Althahban, N. Dimitratos, Q. He, D. J. Morgan, R. L. Jenkins, D. J. Willock and S. H. Taylor, *Science*, 2017, **358**, 223–227.
72 H. Du, X. Li, Z. Cao, S. Zhang, W. Yu, F. Sun, S. Wang, J. Zhao, J. Wang and Y. Bai, *Appl. Catal., B*, 2023, **324**, 122291.
73 L. Luo, Z. Gong, Y. Xu, J. Ma, H. Liu, J. Xing and J. Tang, *J. Am. Chem. Soc.*, 2021, **144**, 740–750.
74 Y. Sun, S. Liu, H. Chang, J. Liu and L. Piao, *ACS Appl. Mater. Interfaces*, 2025, **17**, 15347–15356.
75 H. Chen and A. A. Adesina, *J. Chem. Technol. Biotechnol.*, 1994, **60**, 103–113.
76 J. T. Niu, Y. L. Wang, Y. Y. Qi, A. H. Dam, H. M. Wang, Y. A. Zhu, A. Holmen, J. Y. Ran and D. Chen, *Fuel*, 2020, 266.
77 L.-l. Xu, H. Wen, X. Jin, Q.-m. Bing and J.-y. Liu, *Appl. Surf. Sci.*, 2018, **443**, 515–524.
78 N. Köpfle, K. Ploner, P. Lackner, T. Götsch, C. Thurner, E. Carbonio, M. Hävecker, A. Knop-Gericke, L. Schlicker, A. Doran, D. Kober, A. Gurlo, M. Willinger, S. Penner, M. Schmid and B. Klötzer, *Catalysts*, 2020, **10**, 1000.
79 X. Zhai, Y. Cheng, Z. Zhang, Y. Jin and Y. Cheng, *Int. J. Hydrogen Energy*, 2011, **36**, 7105–7113.
80 B. Y. Lee, S.-C. Jang, J. Han, V. Cigolotti, S.-H. Choi and S. W. Nam, *Sci. Adv. Mater.*, 2019, **11**, 629–641.