npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01664-9

Machine learning accelerated descriptor design for catalyst discovery in CO_2 to methanol conversion

Prajwal Pisal [1,2,3,6], Ondřej Krejčí [1,4,6] & Patrick Rinke [1,2,3,5]

Transforming CO_2 into methanol represents a crucial step towards closing the carbon cycle, with thermoreduction technology nearing industrial application. However, obtaining high methanol yields and ensuring the stability of heterocatalysts remain significant challenges. Herein, we present a sophisticated computational framework to accelerate the discovery of thermal heterogeneous catalysts, using machine-learned force fields. We propose a new catalytic descriptor, termed adsorption energy distribution, that aggregates the binding energies for different catalyst facets, binding sites, and adsorbates. The descriptor is versatile and can be adjusted to a specific reaction through careful choice of the key-step reactants and reaction intermediates. By applying unsupervised machine learning and statistical analysis to a dataset comprising nearly 160 metallic alloys, we offer a powerful tool for catalyst discovery. We propose new promising candidates such as ZnRh and $ZnPt_3$, which to our knowledge, have not yet been tested, and discuss their possible advantage in terms of stability.

Utilizing CO_2 in the production of useful chemicals closes the carbon loop and subsequently reduces CO_2 emissions. Converting CO_2 into liquid fuels or chemical feedstocks like methanol can decrease our dependence on fossil fuels[1]. The hydrogenation of CO_2 to methanol involves the reaction of two gases, similar to other important chemical processes like the Haber-Bosch synthesis[2], which produces ammonia, a precursor for fertilizers, from hydrogen and nitrogen. Both processes occur in thermochemical reactors and face significant energetic barriers, requiring high temperatures and pressures to yield the desired products. Heterogeneous catalysis is a key method to lower these reaction barriers, making the processes both technologically and economically viable[2,3]. However, the economic feasibility of methanol synthesis has not yet been achieved[4].

The identification of an ideal CO_2 conversion method focuses primarily on two technological pathways: thermochemical[5] and electrochemical[6]. The thermochemical approach offers significant potential for rapid industrial adoption due to its resemblance to syngas conversion[5]. Current catalysts, typically based on the industrial syngas catalyst $Cu/ZnO/Al_2O_3$, suffer from low conversion rates, low selectivity[7], and oxidation poisoning[8]. Addressing these issues with better catalysts could increase performance and reduce costs[4]. However, experimentally screening materials to discover effective catalysts remains challenging due to the slow and expensive nature of catalyst testing and the vastness of the materials space.

Computational methods such as density functional theory (DFT) could provide a complementary, efficient and cost-effective alternative for catalyst discovery. However, calculating turn-over frequencies based on reaction barriers is often computationally intensive because it requires explicit transition state calculations. Moreover, many other factors like catalyst microstructure, reactor designs, mass and heat flows, etc. affect catalytic performance and necessitate multi-scale modeling[9,10]. Consequently, approximate methods and concepts, such as the Sabatier principle that relate catalytic activity to the adsorption energies of reaction intermediates calculated using DFT[11], have been frequently employed in extensive searches for candidate materials[11–13]. Over the years, numerous approximations have been developed that have guided catalyst search, extending the Sabatier principle to correlate activity with more easily obtainable activity descriptors, such as *d*-band center and scaling relations[12,14,15]. While these descriptors have provided valuable insight, their usefulness is often constrained to certain surface facets of material or a limited number of material families, such as *d*-metals.

[1]Department of Applied Physics, Aalto University, Espoo, Finland. [2]Department of Physics, Technical University of Munich, Garching, Germany. [3]Munich Center for Machine Learning (MCML), Munich, Germany. [4]Department of Mechanical and Materials Engineering, University of Turku, Turku, Finland. [5]Atomistic Modeling Center, Munich Data Science Institute, Technical University of Munich, Garching, Germany. [6]These authors contributed equally: Prajwal Pisal, Ondřej Krejčí
e-mail: patrick.rinke@aalto.fi

Machine learning (ML) has recently emerged as a powerful alternative in materials research and is gaining traction, with several disciplines employing these techniques to expedite the discovery of new materials[13,16–23]. Data-driven algorithms can analyze vast datasets of catalyst properties and performance, identifying complex relationships that may be beyond the reach of traditional descriptors[24,25]. In the realm of heterogeneous catalysis, the ML methods can primarily be divided into two categories: mapping catalyst activity using new approximate descriptors[18,20,22,26] and prediction of adsorption energies using machine-learned force fields (MLFF)[17,21].

Both of these approaches have significantly boosted effectiveness in heterogeneous catalyst research[27]. Descriptor-based approaches such as SISSO[18] and the latest generation of MLFFs based on graph neural networks[21], offer remarkable precision. However, these methods require training, which results in a computational cost comparable to direct DFT calculations.

In contrast, pre-trained MLFFs, that make use of extensive DFT datasets, offer explicit relaxation of adsorbates on catalyst surfaces and a significant speed-up (a factor of 10^4 or more) compared to DFT calculations while maintaining quantum mechanical accuracy[28,29] making them suitable for high-throughput catalyst discovery workflows. Modern Sabatier principle-based approaches also utilize MLFFs or specialized models to find (global) minimum adsorption energies across multiple material facets[13,21] providing a scalable and transferable framework for screening a broad range of catalysts. Although several studies emphasize the importance of various catalyst facets[9,30], these approaches predominantly utilize data from individual facets for material characterization. Consequently, the challenge persists: how can we effectively predict catalytic performance without limiting our scope to specific material families or facet orientations?

The absence of an adequate descriptor for the activity of complex materials motivates our exploration of methods to better represent contemporary industrial catalysts. These catalysts, composed of nanostructures with diverse surface facets and adsorption sites, present significant challenges to understanding their performance. This work seeks to address these challenges by focusing on three critical objectives: (1) developing a novel descriptor that captures the structural and energetic complexity of catalysts, (2) establishing an efficient workflow for large-scale computational screening, and (3) devising a robust framework to identify promising candidates from the resulting data.

Firstly, we seek to define a descriptor that encapsulates the inherent complexity of heterocatalytic materials. In this study, we introduce adsorption energy distributions (AEDs) as a tool to represent the spectrum of adsorption energies across various facets and binding sites of nanoparticle catalysts. Building on recent advances in characterizing structurally complex materials, such as high-entropy alloys[31,32], we explore the potential of AEDs to fingerprint the material catalytic properties, using the CO_2 to methanol conversion reaction as a case study.

Secondly, we aim to establish a high-throughput and ML-enhanced workflow to accelerate the screening of catalytic materials using our newly formulated descriptor. Traditional density functional theory (DFT) approaches are computationally prohibitive for large-scale studies. To overcome this limitation, we leverage MLFFs from the Open Catalyst Project (OCP)[33,34], enabling rapid and accurate computation of adsorption energies. Our workflow generates an extensive dataset of AEDs, capturing over 877,000 adsorption energies across nearly 160 materials relevant to the CO_2 to methanol conversion reaction. To target the enhanced reliability of our prediction and the effective use of ML models in catalyst discovery, we design a robust validation protocol.

Finally, we address the objective of developing a method to compare AED descriptors, which encode the energy landscape of materials, to those of known effective catalysts. This involves employing unsupervised learning techniques to analyze the extensive dataset of AEDs generated in this study. By treating AEDs as probability distributions, we quantify their similarity using the Wasserstein distance metric[35] and perform hierarchical clustering to group catalysts with similar AED profiles. This approach enables us to systematically compare the AEDs of new materials to those of established catalysts, identifying potential similarities that suggest comparable performance. Through this comparison with known effective catalysts, we seek to identify new materials with similar AEDs, highlighting a few promising candidates for further investigation.

Results and discussion

In order to discover potential new catalysts for converting CO_2 into methanol, we present the workflow depicted in Fig. 1. The key steps are summarized here, with detailed implementation procedures and configurations available in the Methods section.

Search space selection

To effectively reduce the search space for potential catalyst materials for CO_2 thermal conversion, we first isolated the metallic elements that have undergone prior experimentation for this process, as documented by

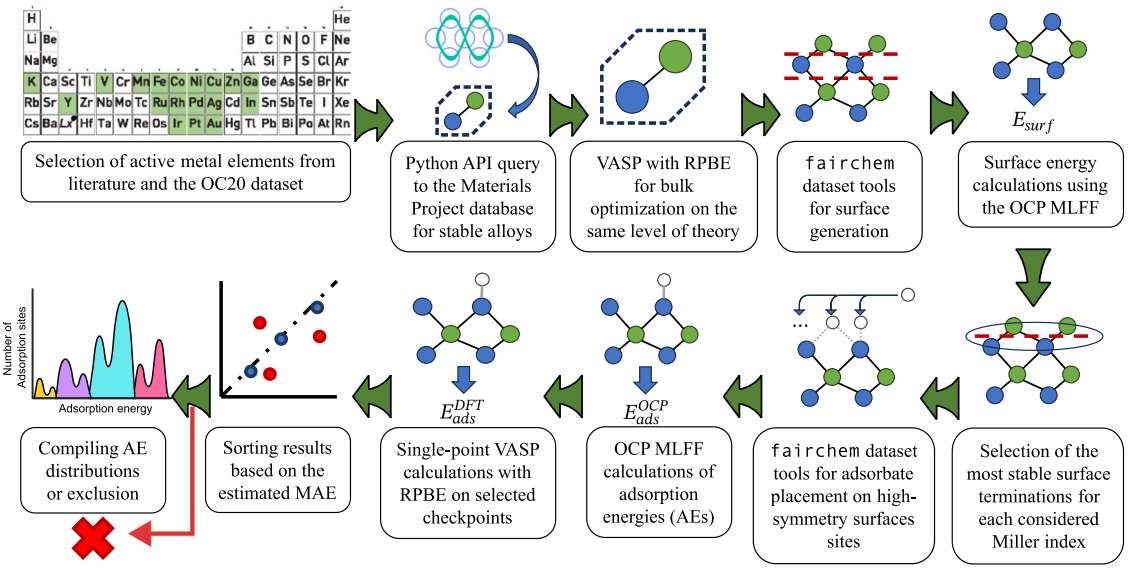

Fig. 1 | Schematic of the workflow for adsorption energy distributions (AEDs) generation. The AED catalyst database was created through a series of steps, including the choice of metals, bulk optimization, selection of relevant surface geometries, preparation of adsorbate geometries, validation and compilation of AEDs, as illustrated in the figure.

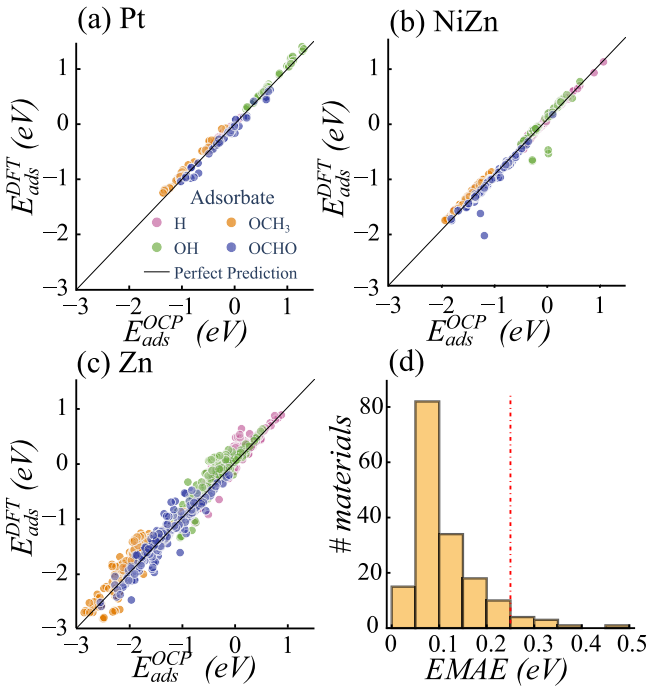

Fig. 2 | Validation of results. a–c Comparison of adsorption energies predicted by the OCP `equiformer_V2` MLFF against single-point DFT calculations, for **a** Pt, **b** Zn and **c** NiZn. **d** Histogram of EMAEs for materials calculated with the OCP MLFF, with a 0.25 eV cut-off line, showing that the majority of materials have their EMAEs between 0.05 and 0.10 eV. Out of a total of 187 OCP-calculated materials, 17 are not shown here as their EMAEs exceed 0.5 eV.

Table 1 | Mean absolute error (MAE) between OCP MLFF predictions and single-point DFT calculations for three selected materials – Pt, Zn and NiZn alloy

Material:	**Pt**				
Adsorbate	***H**	***OH**	***OCHO**	***OCH_3**	**Overall**
MAE (eV)	0.02	0.05	0.06	0.09	**0.06**
EMAE (eV)	0.02	0.07	0.04	0.10	**0.06**
MAE/EMAE	0.97	0.66	1.59	0.93	**0.96**
Material:	**Zn**				
Adsorbate	***H**	***OH**	***OCHO**	***OCH_3**	**Overall**
MAE (eV)	0.10	0.13	0.10	0.15	**0.12**
EMAE (eV)	0.06	0.10	0.06	0.08	**0.07**
MAE/EMAE	1.68	1.34	1.68	1.95	**1.64**
Material:	**NiZn**				
Adsorbate	***H**	***OH**	***OCHO**	***OCH_3**	**Overall**
MAE (eV)	0.02	0.07	0.09	0.06	**0.06**
EMAE (eV)	0.02	0.05	0.09	0.02	**0.05**
MAE/EMAE	0.73	1.36	1.04	2.61	**1.27**

The MAE is compared with the estimated MAE (EMAE), which is obtained from single-point DFT calculations for only three selected structures per adsorbate-material combination. The overall values (shown in bold) represent performance across all adsorbates combined.

Bahri et al.[36]. To maintain prediction accuracy, these elements also had to be part of the Open Catalyst 2020 (OC20) database[33]. The elements shortlisted are the following: K, V, Mn, Fe, Co, Ni, Cu, Zn, Ga, Y, Ru, Rh, Pd, Ag, In, Ir, Pt, and Au. We then proceeded to search through the Materials Project database[37] for stable and experimentally observed crystal structures associated with these metals and their bimetallic alloys. We compiled 216 stable phase forms involving both single metals and bimetallic alloys corresponding to our set of 18 elements. A detailed listing of these materials is provided in Supplementary Tables S1 and S2 of the Supplementary Information. We performed bulk DFT optimization at the RPBE[38] level to align with OC20 for the obtained materials. Optimization of 22 materials was not successful, and therefore, they were excluded from the materials list, as detailed in Supplementary Table S2 in the Supplementary Information.

To identify the most crucial adsorbates for AEDs calculations, we perused the existing literature. An experimental investigation by Amman et al.[30] highlighted the presence of surface-bound radicals such as *H (hydrogen atom), *OH (hydroxy group), *OCHO (formate), and *OCH_3 (methoxy) as essential reaction intermediates in the thermocatalytic reduction of CO_2 to methanol. Based on these findings, we selected these adsorbates for our AEDs calculations. Please note that the notation for formate can vary across the literature, e.g., *HCOO[30,39] and HCOO*[40]. With the help of `fairchem` repository tools by OCP[41], we created surfaces with their Miller index $\in \{-2, -1, \ldots, 2\}$ and calculated their total energy using OCP MLFF. If we encountered multiple cuts for the same facet, we selected the one with lowest energy for further calculations. Then we engineered surface-adsorbate configurations for the most stable surface terminations across all facets within our defined Miller index range for the materials, as described in the Methods section, and optimized these configurations using the OCP MLFF. During this process, we discovered that seven materials exhibited so large surface-adsorbate supercells, that their calculations were infeasible on available GPU resources, even with the effective OCP MLFF. Consequently, they were excluded from our study.

Validation and data cleaning

In our work, we have employed the OCP `equiformer_V2` MLFF. Its reported accuracy for the adsorption energy of small molecular fragments is 0.23 eV[42]. However, *OCHO was not included in the OC20 database used for training the `equiformer_V2`, raising concerns about the accuracy of our adsorption energy predictions in this work. To benchmark `equiformer_V2` for our use case, we chose Pt, Zn, and NiZn and performed explicit DFT calculations (see Methods section for details). The comparison between predicted and DFT-calculated adsorption energies can be found in Fig. 2 and Table 1. The predictions for Pt are precise, whereas the NiZn results show some outliers, and there is a noticeable degree of scatter for Zn. Despite this, the overall mean absolute error (MAE) for the adsorption energies of the selected materials is 0.16 eV, which is impressive and falls within the reported accuracy of the employed MLFF.

To affirm the reliability of our predicted AEDs across a broader range of materials along with maintaining computational practicality, we integrated a validation step within our analysis workflow. We sampled the minimum, maximum, and median adsorption energies for each adsorbate-material pair from the predicted AEDs. We performed single-point DFT calculations on these selected systems and compared with the adsorption energy predictions of the OCP MLFF. The difference is compiled in an 'estimated MAE' (EMAE). Comparisons between EMAE and the all-encompassing MAE for our complete test set are presented in Table 1. While the EMAE may differ from the actual MAE by up to a factor of three for specific adsorbates, it generally remains in close proximity to the actual MAE, thus serving as a reliable gauge of data quality.

The validation step is connected with the final data cleaning when we exclude any material with an EMAE surpassing the threshold of 0.25 eV. Consequently, 29 materials were expunged from our dataset, retaining 158 materials. Most materials flagged for significant EMAEs exhibited magnetic properties, exemplified by materials like MnCo, MnGa, or FeCo. Magnetism presents significant challenges for the non-spin-polarized DFT calculations used in OC20 and in this work. A complete list of estimated MAEs for the remaining 158 materials is accessible in ref. 43.

Adsorption energy distributions

Lastly, to compile the AEDs, we examined the relaxed configurations. For many distinct initial configurations of identical adsorbates, materials, and facets that converged to the same final structure, only one of them is considered in the AED. In our final compilation, we transformed all AEDs into

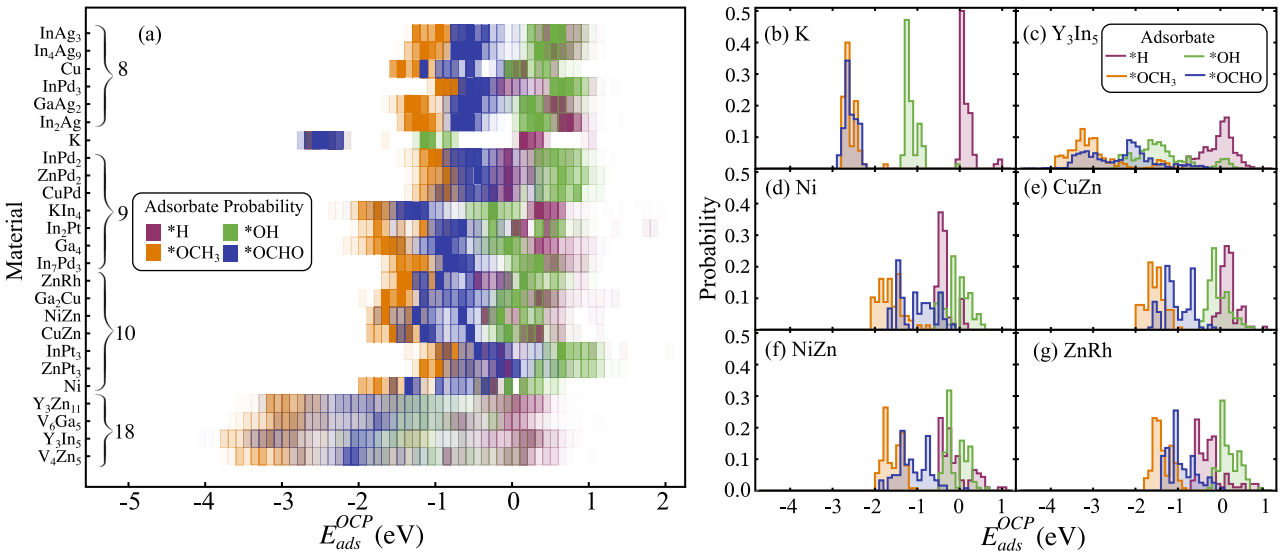

Fig. 3 | AEDs for selected subset of materials. a AEDs for 25 materials with the amplitude of the distribution encoded in the intensity of the corresponding color; detailed AEDs for **b** K, **c** Y_3In_5, **d** Ni, **e** CuZn, **f** NiZn, **g** ZnRh. The numbers on the *y*-axis of panel (**a**) indicate the cluster numbers defined in Fig. 4.

histograms that depict the probability distribution of adsorption sites falling within 0.1 eV energy intervals. Each AED was normalized, ensuring that the aggregate probability of adsorption sites per adsorbate and material equaled one. This standardization facilitates direct comparisons across materials with different numbers of adsorption sites, which can range from several tens to nearly 10,000 for a single material, depending upon the complexity and symmetry of its bulk structure. For illustrative purposes, Fig. 3 displays examples of AEDs for selected materials. The AEDs for all investigated materials is shown in Supplementary Fig. S1 in the Supplementary Information.

Inspection of Fig. 3 and Supplementary Fig. S1 reveals that adsorption energies span a wide interval from −7.42 eV to 2.40 eV. We included energies above zero, although positive adsorption energies are typically indicative of molecular desorption. However, the adsorption energies reported in this work do not include entropy and pressure terms, which could shift the energies to more negative values. Secondly, the adsorption energy of radicals is somewhat ill-defined, if different desorption channels are conceivable. Since our objective is to achieve a qualitative comparison across materials, the price energy zero is of no relevance, as long as it is chosen consistently.

The AEDs exhibit varying dispersion and forms, indicating fluctuations in adsorption energy and related activity levels across the material space. The adsorption energies of $*OCH_3$ (E_{ads}) are generally the lowest, followed by those of *OCHO, which are approximately 0.5–1 eV higher. However, certain materials, such as K (illustrated in Fig. 3b), show unique distribution overlaps for *OCHO and $*OCH_3$. Meanwhile, *H and *OH have comparatively higher E_{ads} values, although their order is inconsistent. For instance, in some cases, *H has the highest E_{ads}, particularly for K and Y_3In_5, whereas the opposite trend is observed for other materials like Ni. Single metal distributions are generally narrower and higher, as seen in the examples of K and Ni. Similarly, alloys composed of elements with high symmetry, such as CuZn, also exhibit narrow AEDs.

If the AEDs of a material predominantly align around the adsorption energy linked to maximal activity according to the Sabatier principle, the material is a strong candidate for a good catalyst. Conversely, complex alloys with low symmetry, such as Y_3In_5 (shown in the lower section of Fig. 3a and in Fig. 3c), display broad AED spreads. Extremely low adsorption energies can lead to catalyst poisoning, while excessively high energies can significantly reduce catalytic activity. Therefore, broad distributions are less desirable, as only a small portion of the material's surface contributes effectively to catalytic processes.

Unsupervised learning: catalyst discovery

Although the ideal AEDs for the four adsorbates remain unknown, it is feasible to approximate their reactivity using AEDs based on their resemblance to previously identified, efficacious catalytic materials. In this context, our AEDs can be conceptualized as four-dimensional probability distributions. To quantify similarities across AEDs of different materials, we employ the Wasserstein distance as the metric[35]. By computing Wasserstein distances for all possible material combinations, we construct a distance matrix. To interpret the distance matrix, we apply hierarchical agglomerative clustering with Ward linkage[44], which facilitates the identification of materials with similar AEDs. The outcomes of this clustering analysis are depicted in Fig. 4.

For a clustering threshold corresponding to a Wasserstein distance of 2.5×10^{-3}, we arrive at a total of 19 distinct clusters, with potassium (K) forming its own, isolated, unnumbered cluster. The separation between clusters 11–19 and clusters 1–10 is considerable. The distinguishing feature is the broadness of the AEDs. The distributions in clusters 11–19 are noticeably broader than in clusters 1–10. Representative examples are depicted in Fig. 3a. The four materials at the bottom of the figure (Y_3Zn_{11}, V_6Ga_5, Y_3In_5, V_4Zn_5) pertain to cluster 18, whereas the rest belong to clusters 8–10. Further details are available in Supplementary Fig. S1 of the Supplementary Information, presenting the clustering of all considered materials. AEDs exhibit variability across distinct clusters (1–10) but show remarkable similarity within each individual cluster. For example, the AEDs for Ni, CuZn, NiZn, and ZnRh illustrated in Fig. 3d–g belong to the same cluster.

Clusters 8 through 10 aggregate into a larger cluster, from hereon denoted the macro-cluster, with relatively homogeneous AEDs. It encompasses materials such as Cu, a notably active component within known $Cu/ZnO/Al_2O_3$ catalysts[30,36]. The clusters also contain non-Cu materials such as Zn-Pd, Pd-In, Pt-In, and Ni-Zn in different compositions that have been reported as catalytic converters of CO_2 to methanol[36,45]. Also different compositions of the bimetallic alloys Ga-Ag, In-Ag, K-In, Zn-Rh, and Zn-Pt, which, to our knowledge, have not been tested for CO_2 to methanol conversion, are grouped with these materials. While most of the materials in the macro-cluster have either shown good catalytic performance or some have not been tested, potassium (K) (the lone non-numbered cluster), as a pure metal, is likely to undergo rapid oxidation under reaction conditions. Therefore, we anticipate that the macro-cluster, consisting of clusters 8-10, is likely too diverse to pinpoint only catalytically active materials.

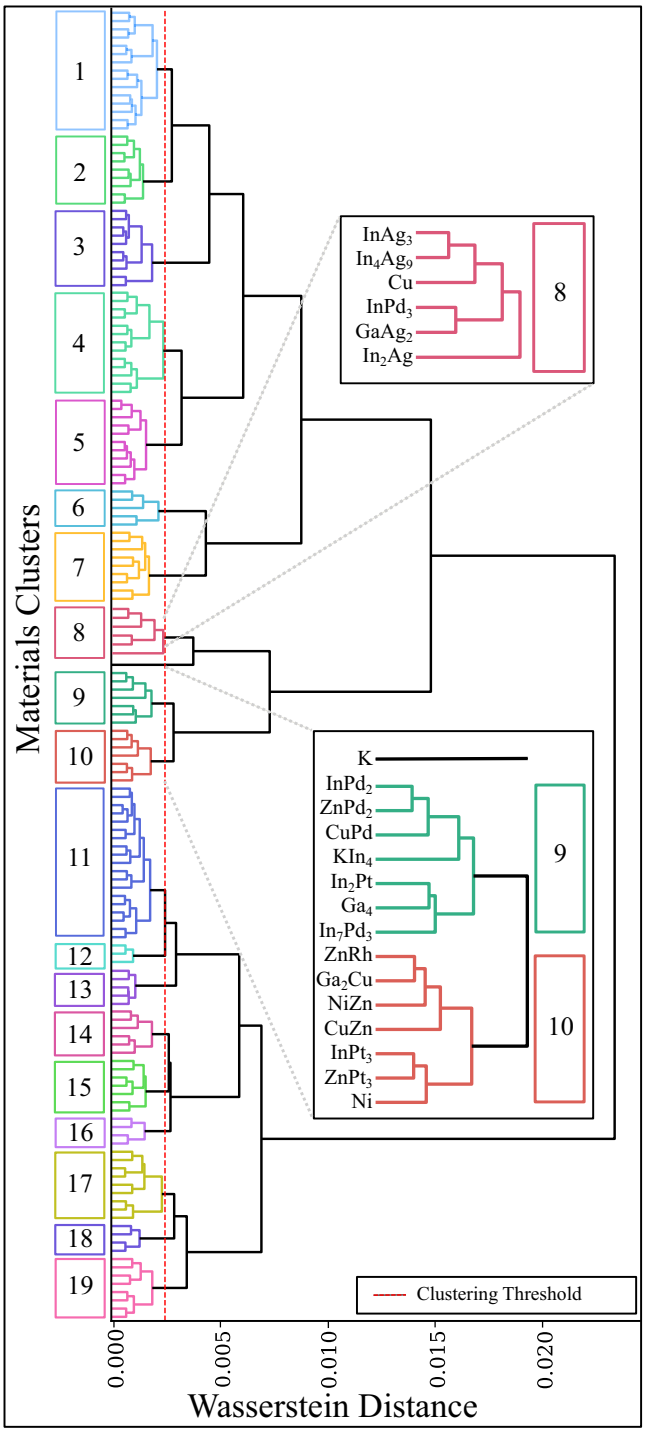

Fig. 4 | Hierarchical clustering of the materials, based on the Wasserstein distances. The graph shows that all 158 materials were assembled into 19 clusters, based on the similarity between of the AEDs, with the exception of potassium, which is dissimilar to nearby materials and forms a single, non-numbered cluster. This can be seen in detail in the insets, where materials in clusters 8, 9 and 10 are shown. The cluster 10 contains several alloys, which are part of known high-yield catalysts as well as new potentially active materials.

Upon closer inspection, ZnRh and $ZnPt_3$ stand out as new candidates. They are part of cluster 10, which also includes Ga_2Cu, NiZn, $InPt_3$, Ni and mainly CuZn, but have not been tested for CO_2 to methanol conversion. While the exact material composition of the Cu and ZnO-based catalysts during the proceeding reaction is still debated[46–48], some studies suggest that the formation of a Cu-Zn alloy enhances the activity. This Cu-Zn alloy, which is part of cluster 10, is believed to contribute to increased activity[30,47,49]. Similarly, NiZn has also been identified as an effective CO_2 catalyst[45]. Catalysts such as $Cu/Ga/ZnO$[50], $Cu/Ga/SiO_2$[51] and Pt/In_2O_3[52], known for their high methanol yield, may include Ga_2Cu and $InPt_3$ alloys, respectively. Finally, Ni is often part of catalysts for CO_2 transformation to methane[53,54]. The strong catalytic activity of Ga_2Cu, NiZn, $InPt_3$, Ni, and mainly CuZn in this cluster suggest that ZnRh and $ZnPt_3$ should also have high activity.

To conclude this section, we reiterate that our approach groups catalyst materials according to their computed AED similarity. To ascribe meaning to certain similarities we observe, we currently rely on experimental reports of catalytically active materials. Our proposals for interesting candidates are based on the assumption that AED similarity with a known good catalyst is a meaningful indicator for promising catalytic activity. Since the catalyst composition and microstructure are often not reported or not known, "active sites" or details of the catalytic mechanisms also remain opaque[30,46,48,55]. In this context, our AED descriptor remains an attempt to find proxies for complex processes. It goes beyond the common practice of focusing on single adsorption energies in "active sites", but could certainly be extended in future work and in collaboration with more detailed experimental investigations.

Statistical analysis and discussion

AEDs could serve as a descriptor of activity, however, the vast number of parameters (at least 388 bins in the distribution) makes it challenging to analyze them manually. To further our insight into the generated data, we conducted a statistical analysis of AEDs (SAAEDs) that facilitates comparison with previous adsorption energy-based studies. An example can be seen in Table 2, where we present the minimum adsorption energies for a subset of materials featured in Fig. 3a.

Our SAAED analysis revisits individual binding energies and connects to the Sabatier principle. For instance, the results for *OH, *OCHO, and $*OCH_3$ can be compared to the volcano plot in Studt et al.[49], that relates the catalytic activity of the studied materials to the oxygen adsorption energy. In line with our approach, their work compares potential catalyst materials to Cu, although their focus lies on single-facet surfaces. Following previous findings that the Cu(211) facet is more active than the close-packed Cu(111) surface[47], Studt et al. use Cu(211) as their reference. The catalytic activity of Cu is further enhanced when Zn is added to the Cu(211) surface (referred to as Cu+Zn in the article). The oxygen adsorption energy decreases upon Zn addition, which indicates that the optimal oxygen adsorption energy should be lower than its minimal adsorption energy on the Cu(211) surface. Our data is consistent with those findings for the majority of our promising candidate materials. The minimal adsorption energy (E_{ads}^{min}) for all the oxygen-containing adsorbates on the majority of the materials in cluster 10 (highlighted in Table 2), including ZnRh, lies below that of Cu (our Cu data also covers the (211) surface) and is closely aligned across the materials. The exceptions are $InPt_3$ and $ZnPt_3$, in which the minima lie slightly above those of Cu, while both materials exhibit similar E_{ads}^{min} for all other adsorbates. This difference suggests that $InPt_3$ and $ZnPt_3$ may feature slightly different CO_2 conversion mechanisms.

Using minimum adsorption energies derived from ML models is comparable to previously studied methods for identifying global minima[13,21]. Although the techniques by Lan et al.[21] and Chen et al.[13] might be more appropriate for the straightforward application of the Sabatier principle, our approach excels in providing more comprehensive information on various facets of catalytic materials. We have compiled this information for selected materials in Table S1 in the Supplementary Information. For example, the AED spread across energies, which can be deduced from the standard deviation E_{ads}^{std}, provides information about the percentage of the surface area usable for catalytic conversion.

Ultimately, both AEDs and SAAEDs, available on Zenodo[43], can serve as material fingerprints. The SAAED acts like a materials descriptor, similar to the Magpie descriptor[16], but can be adapted to specific reactions through the choice of adsorbates, offering more detailed and relevant material

Table 2 | Extract of the statistical analysis on the AEDs

Material	$*H\ E_{ads}^{min}$	$*OH\ E_{ads}^{min}$	$*OCH_3\ E_{ads}^{min}$	$*OCHO\ E_{ads}^{min}$
$InAg_3$	0.24	0.09	−0.88	−1.41
In_4Ag_9	0.12	−0.07	−0.93	−1.53
Cu	−0.14	−0.18	−1.29	−1.64
$InPd_3$	−0.54	0.27	−1.04	−1.22
$GaAg_2$	0.22	−0.25	−1.06	−1.80
In_2Ag	0.27	−0.06	−0.96	−1.59
K	−0.01	−1.31	−2.81	−2.77
$InPd_2$	−0.39	−0.04	−1.09	−1.57
$ZnPd_2$	−0.38	0.06	−1.11	−1.44
CuPd	−0.44	0.03	−1.12	−1.41
KIn_4	−0.40	−1.11	−2.36	−2.46
In_2Pt	−0.33	−0.54	−1.69	−2.12
Ga_4	−0.71	−0.98	−1.73	−2.36
In_7Pd_3	−0.36	−0.33	−1.20	−1.80
ZnRh	−0.63	−0.30	−1.39	−1.71
Ga_2Cu	−0.20	−0.92	−1.36	−2.06
NiZn	−0.49	−0.49	−1.81	−1.93
CuZn	−0.19	−0.44	−1.67	−1.94
$InPt_3$	−0.59	0.13	−1.09	−1.41
$ZnPt_3$	−0.43	0.10	−0.98	−1.40
Ni	−0.53	−0.54	−1.66	−2.05
Y_3Zn_{11}	−0.89	−2.22	−3.92	−3.68
V_6Ga_5	−1.30	−3.37	−5.44	−3.78
Y_3In_5	−0.88	−2.51	−4.61	−3.93
V_4Zn_5	−1.13	−2.29	−3.72	−3.87

The minimum of the OCP MLFF predicted adsorption energies E_{ads}^{min} which is basically comparable to the adsorption energy used in Sabatier principle. We show the predicted E_{ads}^{min} for the same materials as in Fig. 3 (i.e. clusters 8–10 and 18), for all the considered adsorbates. The materials of cluster 10 are highlighted in bold.

information. Optionally, specific descriptors (AED, SAAED) and general descriptors (like Magpie[16]) may be combined to enhance the information that might be lacking in ML models from theoretical calculations.

Both catalyst descriptors are tailored to perform extensive searches for catalytically active candidates. They do not, however, include effects of the support, additives, preparation procedures and operando states that could change the morphology of the catalyst (e.g., nanoparticle sizes or areas of different facets). Our proposed AED descriptor does not take the facet area into account and is therefore insensitive to morphology changes of the catalysts under reaction conditions. In principle, a Wulff construction could introduce better facet information. However, it also cannot account for support effects, additives or preparation conditions.

Our methodology facilitates high-throughput screening of metallic catalyst candidates. At present, the effects of co-operating oxides such as ZnO, In_2O_3, and ZrO_2 that have been observed experimentally are not considered. Such oxides affect the electronic structure and adsorption energy landscape at the metal-oxide interface[46,55–58] and should be included in future versions of our descriptor. For instance, incorporating general descriptors (e.g., Magpie) for co-catalysts and support materials could provide additional information to decide which active material-support combination should be investigated further in experimental testing.

Additionally, the choice of adsorbates can also influence the effectiveness of the AED descriptor. Our study focuses on the four most relevant intermediates, observed on the Cu(211) surface[30]. Studies on different materials, such as Ni-ZrO_2[58,59], suggest that other intermediates or by-products like CO could play an important role in the hydrogenation mechanism. Thus, further investigations could extend the set of adsorbates to better capture various reaction paths and, therefore material-specific activity.

Our workflow clusters materials with high CO_2 conversion efficiency, but the materials can vary in their selectivity towards methanol or methane[36,45,50,52,60]. As the reaction conditions, preparation procedures, or the interaction with support materials seem to affect the selectivity[45,60], our proposed catalyst candidates should thus be tested under various conditions to investigate their optimal selectivity towards methanol.

To finalize the analysis of our results, the similarity of the SAAED and Wasserstein distances of ZnRh and $ZnPt_3$ to good catalysts in the literature suggests that they could be potential catalyst candidates. As Cu-based catalysts are known for their vulnerability to degradation[5], it is therefore reasonable to pre-examine these materials also in terms of stability. Given the harsh reaction conditions, mainly temperatures around 800 K[36], the melting temperature of the catalyst is directly related to the stability of the catalyst. The melting temperature of both ZnRh and $ZnPt_3$ is higher than that of pure copper or CuZn[37], suggesting that our new candidates could also be more durable.

Summary

In summary, we have established a fast and reliable computational approach for discovering new catalyst candidates for the conversion of CO_2 to methanol utilizing data-driven methodologies such as MLFFs and hierarchical clustering. Beginning with a list of potential metallic elements, we extracted experimentally verified materials from the Materials Project database. By integrating tools from `fairchem`, mainly OCP MLFFs, we created an extensive database of adsorption energies for a wide range of material facets and possible adsorption sites. We compiled this information to obtain a novel material descriptor, AED, which offers a more effective representation of the complex nature of heterocatalysts compared to standard methods. By carefully choosing the adsorbates, the descriptor can be tailored to provide the most information for any heterocatalytic reaction under study. Through efficient sampling for validation, we were able to quantify the quality of our workflow with a minimal number of DFT calculations while ensuring the high quality of our database. We grouped the materials by their AED similarity using statistical methods and clustering. This allowed us to pinpoint promising new candidates, namely ZnRh and $ZnPt_3$, based on their resemblance to known effective catalysts. Our results indicate that AEDs, together with statistical analysis, can serve as material fingerprints, aiding in the prediction of catalyst activity and accelerating the discovery process.

Methods

Bulk preparation

We sourced the bulk geometries of experimentally observed metals and alloys of select elements from the Materials Project Database[37]. These structures were selected on the basis of their stability, as indicated by cohesive energies located on the convex hull. We optimized the bulk geometries using the RPBE functional[38] as implemented in the Vienna Abinitio Simulation Package (`VASP`)[61,62]. A plane-wave cutoff of 500 eV was used in the first attempt to relax the structures. If the initial run did not converge, we increased the cut-off to 550 eV. We sampled the Brillouin zone with a k-point spacing of 0.17 $Å^{-1}$. The calculations were automated using the workflows developed by Atomate[63] based on the Pymatgen[64], Custodian[64] and Fireworks[65] libraries.

Surface generation and selection

We generated all symmetrically distinct surfaces of Miller indices $\in \{-2, -1, 0, \ldots, 2\}$ from the relaxed bulk structures using the workflow implemented by OCP, i.e. `fairchem`[33,41]. We fixed the thickness of the slabs to 7 Å with a vacuum of 20 Å along the z-direction. These parameters were chosen to be consistent with the OC20 dataset that was used to train the

OCP MLFFs to maintain high prediction accuracy[33]. With the pre-trained `gemnet-oc` MLFF[66], we relaxed the surfaces and obtain their *total energy*. When encountering different surface terminations for a given facet, i.e., different absolute positions of the surface plane, we retained only the structure with the lowest energy.

Generation and relaxation of adsorbate-surface configurations

The selected surfaces were used to generate adsorbate-surface configurations using neutral fragments of *H, *OH, *OCHO and *OCH_3 by means of the `fairchem` input generation workflow[41]. The systems were relaxed using an *adsorption energy*-based OCP `equiformer_V2` equivariant graph neural network MLFF to obtain the relaxed geometries. The corresponding adsorption energy E_{ads} follows the OCP convention $E_{ads} = E_{system} - E_{surface} - E_{adsorbate}$, where the gas phase energy $E_{adsorbate}$ is in this case the energy of radicals, and therefore hard to define. The OCP MLFFs are using the definition of Chanussot et al.[33], which is a linear combination of atomic energies for H, O and C, derived from H_2, H_2O and CO, respectively. Here, $E_{adsorbate}$ is −3.477 eV for *H, −10.552 eV for *OH, −24.917 eV for *OCH_3 and −25.161 eV for *OCHO. As our work is focusing on identifying AED similarities or differences rather than specific adsorbate values, different adsorbate references would merely shift all AEDs by the same value, but will not affect our results. The force convergence criterion for the adsorbate relaxations was set to 0.03 eV/Å. The `equiformer_V2` relaxations were performed using `NVIDIA Ampere A100` GPUs, completing each relaxation in just a few seconds, whereas corresponding DFT calculations would require several hours on two `AMD Rome 7H12` CPUs with 64 cores each. Relaxed configurations of the same material and facet were compared to identify redundancies. If two or more configurations exhibited identical adsorbate geometries (and thus energies) within a tolerance of 0.1 Å in each spatial direction, only one configuration was retained for further consideration. Nonunique configurations and their corresponding adsorption energies were excluded. With this, we ensured that the remaining data represent meaningful adsorption positions and thus capture the material's true energetic landscape.

We aggregated the adsorption energies into AEDs for every adsorbate and material. The AED is represented by a histogram of 0.1 eV wide bins, that are centered along energies $\{-7.4, -7.3, \ldots . 2.4\}$. The histogram settings were chosen so that each AED has at least five empty bins on either side, which is important for further analysis. We normalized the AEDs by dividing the entry in each bin by the number of unique adsorption configurations.

Estimation of the mean absolute error for the adsorption energy prediction

We initially evaluated the performance of the `equiformer_V2` model on Pt, Zn, and NiZn. For this evaluation, we used adsorbate-surface geometries relaxed with `equiformer_V2` and clean surface geometries optimized using `gemnet-oc`. We then computed the DFT adsorption energies without further relaxations. We used the same `VASP` settings as for the generation of the OC20 dataset[33], but increased the plane-wave cutoff to 450 eV. This adjustment was made on the basis of our preliminary estimate for selected materials to ensure high accuracy of both predicted and calculated adsorption energies. The resulting adsorption energies are shown in Fig. 2.

To determine the EMAE, we employed a computationally efficient strategy to validate the predictions of the OCP MLFF model. From the AED of a given material-adsorbate pair, we chose the three configurations that correspond to the mean, median, and maximum adsorption energies. Subsequently, we carried out single-point DFT calculations for the corresponding adsorbate-surface systems and clean surfaces to compute the adsorption energies.

The EMAE was determined as the average of the three absolute errors. As discussed in the Results section, materials with EMAEs exceeding 0.25 eV were excluded from further analysis due to their insufficient accuracy.

Unsupervised learning

We concatenated the four AEDs for each material into a single distribution, where the added buffer ensures no possible overlap between distributions. We then computed Wasserstein distances for all pairs of materials to create a single distance matrix. Given two 1D probability mass functions, μ and ν, the first Wasserstein distance between the distributions is defined as[67]:

$$W_1(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} ||x-y|| d\gamma(x,y) \quad (1)$$

where $\Gamma(\mu, \nu)$ is the set of (probability) distributions on $\mathbb{R}\times\mathbb{R}$ whose marginals are μ and ν on the first and second factors, respectively. For a given value x, $\mu(x)$ gives the probability of μ at position x, and the same for $\nu(x)$.

Using these Wasserstein distances, we performed agglomerative hierarchical clustering with Ward linkage[44,68,69]. We utilized the Python-based SciPy library[67] to compute the distances and perform clustering on the distance matrix. We used a clustering threshold of 0.0025 to define the maximum distance at which clusters are merged.

Data availability

The data published in this study can be found in ref. 43, Pisal, P., Krejci, O.: Final Geometries and Energies, Statistical Analysis and Estimated Errors of Single Metals and Bimetallics for CO_2 to Methanol Conversion. https://doi.org/10.5281/zenodo.15587232.

Received: 4 December 2024; Accepted: 20 May 2025;
Published online: 04 July 2025

References

1. Ye, R.-P. et al. CO_2 hydrogenation to high-value products via heterogeneous catalysis. *Nat. Commun.* **10**, 5698 (2019).
2. Rohr, B. A., Singh, A. R. & Nørskov, J. K. A theoretical explanation of the effect of oxygen poisoning on industrial Haber-Bosch catalysts. *J. Catal.* **372**, 33–38 (2019).
3. Saito, M., Fujitani, T., Takeuchi, M. & Watanabe, T. Development of copper/zinc oxide-based multicomponent catalysts for methanol synthesis from carbon dioxide and hydrogen. *Appl. Catal. A Gen.* **138**, 311–318 (1996).
4. Nyári, J. et al. Techno-economic barriers of an industrial-scale methanol CCU-plant. *J. CO2 Util.* **39**, 101166 (2020).
5. Ganesh, I. Conversion of carbon dioxide into methanol - a potential liquid fuel: Fundamental challenges and opportunities (a review). *Renew. Sustain. Energy Rev.* **31**, 221–257 (2014).
6. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **1**, 696–703 (2018).
7. Wang, L., Etim, U. J., Zhang, C., Amirav, L. & Zhong, Z. CO_2 activation and hydrogenation on Cu-ZnO/Al_2O_3 nanorod catalysts: an in situ FTIR study. *Nanomaterials* **12**, 2527 (2022).
8. Li, D., Wang, Z., Jin, S. & Zhu, M. Deactivation and regeneration of the commercial Cu/ZnO/Al_2O_3 catalyst in low-temperature methanol steam reforming. *Sci. China Chem.* **66**, 3645–3652 (2023).
9. Bruix, A., Margraf, J. T., Andersen, M. & Reuter, K. First-principles-based multiscale modelling of heterogeneous catalysis. *Nat. Catal.* **2**, 659–670 (2019).
10. Posada-Borbón, A. & Grönbeck, H. A first-principles-based microkinetic study of CO_2 reduction to CH_3OH over In_2O_3(110). *ACS Catal.* **11**, 9996–10006 (2021).
11. Che, M. Nobel prize in chemistry 1912 to Sabatier: organic chemistry or catalysis? *Catal. Today* **218-219**, 162–171 (2013).
12. Medford, A. J. et al. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* **328**, 36–42 (2015).

13. Chen, Z. W. et al. Unusual Sabatier principle on high entropy alloy catalysts for hydrogen evolution reactions. *Nat. Commun.* **15**, 359 (2024).
14. Jones, G., Bligaard, T., Abild-Pedersen, F. & Nørskov, J. K. Using scaling relations to understand trends in the catalytic activity of transition metals. *J. Phys. Condens. Matter* **20**, 064239 (2008).
15. Vogt, C. & Weckhuysen, B. M. The concept of active site in heterogeneous catalysis. *Nat. Rev. Chem.* **6**, 89–111 (2022).
16. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Mater.* **2**, 16028 (2016).
17. Ulissi, Z. W. et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. *ACS Catal.* **7**, 6600–6608 (2017).
18. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
19. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
20. Zhang, N. et al. Machine learning in screening high performance electrocatalysts for CO_2 reduction. *Small Methods* **5**, 2100987 (2021).
21. Lan, J., Palizhati, A. & Shuaibi, M. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Comput Mater.* **9**, 172 (2023).
22. Mou, L.-H., Han, T., Smith, P. E. S., Sharman, E. & Jiang, J. Machine learning descriptors for data-driven catalysis study. *Adv. Sci.* **10**, 2301020 (2023).
23. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Computational Mater.* **4**, 25 (2018).
24. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. A bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *npj Computational Mater.* **6**, 177 (2020).
25. Fiedler, L. et al. Predicting electronic structures at any length scale with machine learning. *npj Computational Mater.* **9**, 115 (2023).
26. Lunger, J. R., Karaguesian, J. & Chun, H. Towards atom-level understanding of metal oxide catalysts for the oxygen evolution reaction with machine learning. *npj Computational Mater.* **10**, 80 (2024).
27. Liu, X. & Peng, H.-J. Toward next-generation heterogeneous catalysts: empowering surface reactivity prediction with machine learning. *Engineering* **39**, 25–44 (2024).
28. Kang, P.-L., Shang, C. & Liu, Z.-P. Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. *Acc. Chem. Res.* **53**, 2119–2129 (2020).
29. Chen, D., Shang, C. & Liu, Z.-P. Machine-learning atomic simulation for heterogeneous catalysis. *npj Computational Mater.* **9**, 2 (2023).
30. Amann, P. et al. The state of zinc in methanol synthesis over a Zn/ZnO/Cu(211) model catalyst. *Science* **376**, 603–608 (2022).
31. Batchelor, T. A. et al. High-entropy alloys as a discovery platform for electrocatalysis. *Joule* **3**, 834–845 (2019).
32. Pedersen, J. K., Batchelor, T. A., Bagger, A. & Rossmeisl, J. High-entropy alloys as catalysts for the CO_2 and CO reduction reactions. *ACS Catal.* **10**, 2169–2176 (2020).
33. Chanussot, L. et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
34. Tran, R. et al. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **13**, 3066–3084 (2023).
35. Ramdas, A., García Trillos, N. & Cuturi, M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19**, 47 (2017).
36. Bahri, S., Pathak, S., Singhahluwalia, A., Malav, P. & Upadhyayula, S. Meta-analysis approach for understanding the characteristics of CO_2 reduction catalysts for renewable fuel production. *J. Clean. Prod.* **339**, 130653 (2022).
37. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
38. Hammer, B., Hansen, L. B. & Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B* **59**, 7413–7421 (1999).
39. Alam, M. I. et al. Mechanistic and multiscale aspects of thermo-catalytic CO_2 conversion to C_1 products. *Catal. Sci. Technol.* **11**, 6601–6629 (2021).
40. Wu, Z., Cole, J., Fang, H. L., Qin, M. & He, Z. Revisiting catalyst structure and mechanism in methanol synthesis. *J. Adv. Nanomater.* **2**, 1–10 (2017).
41. FAIR-Chem: fairchem: A FAIR-chem project repository. *GitHub*, https://github.com/FAIR-Chem/fairchem (2024).
42. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. Preprint at https://arxiv.org/abs/2306.12059 (2023).
43. Pisal, P. & Krejci, O. Final geometries and energies, statistical analysis and estimated errors of single metals and bimetallics for CO_2 to methanol conversion. https://doi.org/10.5281/zenodo.15587232.
44. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
45. Dongapure, P. et al. Mechanistic insights into near ambient pressure activity of intermetallic NiZn/TiO_2 catalyst for CO_2 conversion to methanol. *ChemCatChem* **15**, 202201150 (2023).
46. Beck, A., Newton, M. A., Water, L. G. A. & Bokhoven, J. A. The enigma of methanol synthesis by Cu/ZnO/Al_2O_3-based catalysts. *Chem. Rev.* **124**, 4543–4678 (2024).
47. Behrens, M. et al. The active site of methanol synthesis over Cu/ZnO/Al_2O_3 industrial catalysts. *Science* **336**, 893–897 (2012).
48. Laudenschleger, D., Ruland, H. & Muhler, M. Identifying the nature of the active sites in methanol synthesis over Cu/ZnO/Al_2O_3 catalysts. *Nat. Commun.* **11**, 3898 (2020).
49. Studt, F. et al. Discovery of a Ni-Ga catalyst for carbon dioxide reduction to methanol. *Nat. Chem.* **6**, 320 (2014).
50. Toyir, J., Ramirez de la Piscina, P., Fierro, J. L. G. & Homs, N. Catalytic performance for CO_2 conversion to methanol of gallium-promoted copper-based catalysts: influence of metallic precursors. *Appl. Catal. B Environ.* **34**, 255–266 (2001).
51. Medina, J. C. & Karelovic, A. Catalytic consequences of Ga promotion on Cu for CO_2 hydrogenation to methanol. *Catal. Sci. Technol.* **7**, 3375–3387 (2017).
52. Men, Y.-L. et al. Highly dispersed Pt-based catalysts for selective CO_2 hydrogenation to methanol at atmospheric pressure. *Chem. Eng. Sci.* **200**, 167–175 (2019).
53. Tang, X., Song, C. & Li, H. Thermally stable Ni foam-supported inverse $CeAlO_x$/Ni ensemble as an active structured catalyst for CO_2 hydrogenation to methane. *Nat. Commun.* **15**, 3115 (2024).
54. Hu, F., Ye, R., Lu, Z.-H., Zhang, R. & Feng, G. Structure-activity relationship of Ni-based catalysts toward CO_2 methanation: recent advances and future perspectives. *Energy Fuels* **36**, 156–169 (2022).
55. Lempelto, A., Gell, L., Kiljunen, T. & Honkala, K. Exploring CO_2 hydrogenation to methanol at a CuZn-ZrO_2 interface via DFT calculations. *Catal. Sci. Technol.* **13**, 4387–4399 (2023).
56. Wang, J. et al. High-performance M_aZrO_x (M_a = Cd, Ga) solid-solution catalysts for CO_2 hydrogenation to methanol. *ACS Catal.* **9**, 10253–10259 (2019).
57. Cheula, R., Tran, T. A. M. Q. & Andersen, M. Unraveling the effect of dopants in zirconia-based catalysts for CO_2 hydrogenation to methanol. *ACS Catal.* **14**, 13126–13135 (2024).

58. Cannizzaro, F., Hensen, E. J. M. & Filot, I. A. W. The promoting role of Ni on In_2O_3 for CO_2 hydrogenation to methanol. *ACS Catal.* **13**, 1875–1892 (2023).
59. Gao, D., Li, W., Wang, H., Wang, G. & Cai, R. Heterogeneous catalysis for CO_2 conversion into chemicals and fuels. *Trans. Tianjin Univ.* **28**, 245–264 (2022).
60. Liu, L., Gao, Y., Zhang, H., Kosinov, N. & Hensen, E. J. M. Ni and ZrO_2 promotion of In_2O_3 for CO_2 hydrogenation to methanol. *Appl. Catal. B Environ. Energy* **356**, 124210 (2024).
61. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
62. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
63. Mathew, K. et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Mater. Sci.* **139**, 140–152 (2017).
64. Ong, S. P. et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Mater. Sci.* **68**, 314–319 (2013).
65. Jain, A. et al. Fireworks: a dynamic workflow system designed for high-throughput applications. *Concurrency Comput.: Pract. Experience* **27**, 5037–5059 (2015).
66. Gasteiger, J. et al. GemNet-OC: developing graph neural networks for large and diverse molecular simulation datasets. https://arxiv.org/abs/2204.02782 (2022).
67. Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).
68. Müllner, D. Modern hierarchical, agglomerative clustering algorithms https://arxiv.org/abs/1109.2378 (2011).
69. Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**, 22–29 (2001).

Acknowledgements

The authors would like to thank Annukka Santasalo-Aarnio and Arpad Toldy for fruitful discussions. O.K. and P.P. express their gratitude to Kirby Broderick, Adeesh Koluru, Brook Wander, John Kitchin, and other Kitchin research group members at Carnegie Mellon University and Zachary Ulissi at Meta, for their help with the OCP models. This project received funding from the European Union – NextGenerationEU instrument and the Research Council of Finland's AICon project (grant number no. 348179). The authors gratefully acknowledge CSC – IT Center for Science, Finland, and the Aalto Science-IT project for generous computational resources.

Author contributions

P.P. and O.K. created the workflow and proceeded the calculations. P.R. supervised the work. All authors contributed to the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01664-9.

Correspondence and requests for materials should be addressed to Patrick Rinke.