Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d5sc06442b

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

Explainable artificial intelligence for materials discovery: application to catalysts for the HER and ORR

Valentin Vassilev-Galindo †*[a] and Javier LLorca*[ab]

The extraordinary progress of strategies coupling *ab initio* calculations and machine learning (ML) has opened the door for both fast and accurate chemical/physical property predictions and for the virtual design of materials. However, these techniques are very often used as a "black box" with the sole objective of obtaining high accuracy with scarce or no special attention on how ML models obtain their predictions. This can be improved by leveraging explainability of ML models, which, at the same time, would increase the chance of ML to offer new insights into the chemistry and physics of materials. Hence, the next generation of ML models in these realms must guarantee explainability by embedding explainable artificial intelligence (XAI) tools into their pipelines. Specifically, ML-assisted materials discovery and design can take great advantage of the use of XAI. Enabling explanations would increase the impact of these approaches by providing not only a set of candidates, but also insights into what makes a given material better than others. With this in mind, using the example of heterogeneous catalysts for hydrogen production and energy generation, here we propose a novel strategy for materials design based on counterfactual explanations. We were able to find materials featuring properties close to the design targets that were later validated with density functional theory calculations. Explanations were devised by comparing original samples, counterfactuals, and discovered candidates. Such explanations allowed us to unveil subtle relationships between the most relevant features, other, in principle, less important features, and the target property. Since this approach can be applied to different applications, this work provides an alternative to already available designing strategies, such as high-throughput screening or generative models, but that, for the first time, incorporates explainability as its main driving mechanism.

Received 22nd August 2025
Accepted 10th November 2025

DOI: 10.1039/d5sc06442b

rsc.li/chemical-science

1 Introduction

The extraordinary progress of strategies coupling *ab initio* calculations and machine learning (ML) has opened the door for both fast and accurate chemical/physical property predictions and for the virtual design of materials.[1–4] The available literature is populated with a plethora of examples: prediction of electronic properties (*e.g.*, band gap, atomization energy, and adsorption energies),[5–10] modeling of potential-energy surfaces,[11–13] discovery of new materials with desirable properties,[14–16] and designing of drugs with specific targets,[17,18] among others. However, such ML methods and models are very often used as a "black box" with the sole objective of obtaining high accuracy or a desired designed target with scarce or no special attention on how the predictions are obtained. Such a scenario can lead to the so-called "Clever Hans" effect, in which a model gets accurate predictions for the wrong reasons (*e.g.*, by learning spurious correlations). This problem can be mitigated by leveraging interpretability and explainability of ML models.

Interpretability in ML means that the mapping process from input to the output prediction can be understood. Some ML methods used in computational chemistry (*e.g.*, decision trees and their derived methods) are interpretable by design[19–21] and model-specific techniques have been applied to enable some interpretability in more complex ML architectures, such as neural networks.[22,23] Nevertheless, explainability, which refers to the assessment of prediction outcome patterns to understand why a model makes specific decisions, is only partially and not consistently addressed in the field. Indeed, there are only a few investigations where the terms Explainable Artificial Intelligence (XAI) or explainable ML are employed.[23–27] This is striking because, without explainability, one cannot ensure that predictions are obtained through rigorous chemical and

[a]*IMDEA Materials Institute, C/Eric Kandel 2, Getafe, 28906, Madrid, Spain. E-mail: valentin.vassilev@unavarra.es; javier.llorca@imdea.org*

[b]*Department of Materials Science, Polytechnic University of Madrid, E. T. S. de Ingenieros de Caminos, 28040, Madrid, Spain*

† Present address: Department of Statistics, Computer Science and Mathematics, Universidad Pública de Navarra, Pamplona, Spain; Navarra Artificial Intelligence Research Center (NAIR Center), Pamplona, Spain.

physical grounds, and one hinders the possibility for ML models to offer new insights into the chemistry and physics of materials. Hence, the next generation of ML models must guarantee explainability by embedding explainable ML tools into their pipelines.

Currently, mostly the only approach for explainability that is employed is the assessment of feature importances.[28–30] It helps in identifying how each feature contributes to the final prediction. This is an effective approach to gain insights into the chemical/physical concepts and trends learnt by the model, but it is not sufficient. The explanations brought by feature importance analysis are not actionable (*i.e.*, they do not tell how a given input can be changed in order to modify the output). Actionability is a robust attribute of XAI approaches that can lead to valid hypothesis by providing an intuitive understanding of predictions and, thus, can offer unprecedented insights in materials science and chemistry. One example of such an actionable XAI approach is counterfactual explanations. They provide insights into model operation by determining examples or cases that explain the difference between a desired outcome and actual outcome.[31] Despite their potential application, counterfactual approaches are just a recent topic in computational chemistry and there are only very few examples related to molecular and medicinal chemistry in which they have been applied.[32–36] In these studies, counterfactual explanations have been used to unveil through ML predictions in which structural modifications (*e.g.*, addition or removal of functional groups and atomic species) promote or hinder blood–brain barrier permeation, solubility in water, toxicity, HIV activity, or protein kinase inhibitors. Therefore, there is a need for extending the use of counterfactual explanations to other chemical systems and applications.

Specifically, ML-assisted molecular/material discovery and design can take great advantage of the use of counterfactual explanations. Typical strategies, such as generative models and high-throughput screening,[16] have been successfully used to find new potential catalysts,[37,38] drugs,[39,40] and photovoltaics[41] to name a few examples, but they do not offer information on why a given candidate and not others is found to have the desired target property. Enabling those explanations would increase the impact of discovery and design strategies by providing not only a set of candidates, but also insights into what makes a given molecule/material better than others. Also, these insights can implicitly lead to the development of "recipes" of how to modify a material in order to improve or deplete a target property. Motivated by this, here we introduce an XAI strategy for materials design and discovery that uses counterfactual explanations as the cornerstone for finding new candidates with desired properties. Explainability in the designing process is ensured by construction since every candidate found can be linked to the original sample from which the counterfactual was generated. Hence, one can analyze which features and how they changed from the original sample in order to get the sought outcome. We tested our strategy on the example of heterogeneous catalysts of two reactions, the hydrogen evolution reaction (HER) and oxygen reduction reaction (ORR), that are critical for hydrogen production and energy generation, respectively. We were able to find materials featuring properties close to the design targets (the adsorption energies of H, O, and OH on Pt). Density functional theory (DFT) calculations for the designed candidates confirmed our predictions, hence validating the proposed XAI strategy. Moreover, explanations to gain insights into why the discovered surfaces are better than others were devised by comparing original samples, counterfactuals, and discovered candidates. Such explanations allowed us to unveil subtle relationships between the most relevant features, other, in principle, less important features, and the E_{ads}. This approach can be applied to different target properties and materials. For instance, in other fields related to catalysis (*e.g.*, photocatalysis and organometallic chemistry), where adsorption energies also play a relevant role, the application of our approach would be straightforward since the same features could probably be used as descriptors to train the ML models. Therefore, only a suitable target for the generation of counterfactuals would need to be selected. For other applications in materials science, the use of our approach would also require an analysis of the best features to train the models since the target property might be different (*e.g.*, the band gap in photovoltaics) and an adaptation of the steps and filters during the retrieval of candidates that suits the obtained set of features. Hence, this work provides an alternative to already available designing strategies, such as high-throughput screening or generative models, but that for the first time incorporates explainability as its main driving mechanism.

2 Results and discussion

2.1 Example application: adsorption energies in the catalysis of the HER and the ORR

The production of hydrogen and the generation of green energy depend on two key reactions, respectively: the HER, which in acidic media involves the following steps:

$$\begin{gathered} H^+ + * + e^- \rightarrow H^* \\ H^* + H^+ + e^- \rightarrow * + H_2 \\ 2H^* \rightarrow 2* + H_2 \end{gathered} \quad , \tag{1}$$

and the ORR, whose dissociative mechanism in acidic media is composed of

$$\begin{gathered} \frac{1}{2}O_2 + * \rightarrow O^* \\ O^* + H^+ + e^- \rightarrow HO^* \\ HO^* + H^+ + e^- \rightarrow H_2O + * \end{gathered} \quad , \tag{2}$$

where * represents a site on the surface of the catalyst and X* represents the atom/molecule adsorbed.

According to Sabatier's principle,[42] the variation in the E_{ads} of adsorbates involved in a reaction (in this case, H for the HER and O and OH for the ORR) can serve to assess the activity of a catalyst. Namely, it states that the E_{ads} should be neither too high nor too low for reactions passing through an adsorbed intermediate. If the E_{ads} is too high (endothermic), adsorption is slow and limits the overall rate, whereas the catalyst surface becomes poisoned and desorption is limited if the E_{ads} is too low (exothermic). In terms of water splitting electrocatalysis,

this principle leads to the conclusion that the free energy of adsorption should be close to zero at the equilibrium potential.[43,44] Because of this, the magnitude of E_{ads} obtained from DFT calculations can be used to determine the overall rate of the reaction with

$$G_{ads} = E_{ads} + E_{ZPE} - T\Delta S \quad (3)$$

where G_{ads}, E_{ads}, E_{ZPE}, and ΔS stand for the variation of the free and adsorption energies, the zero point energy and the entropy, respectively, during the adsorption of the intermediate species, and T is the absolute temperature. This relationship should be evaluated for each intermediate species that appears in the catalytic process (only H in the HER and O and OH in the ORR), taking into account that the process with the highest free energy barrier will limit the rate of the reaction. Therefore, knowing the E_{ads} of the intermediate species on a material is a first fundamental step to assess a candidate catalyst. An accurate computation of E_{ads} is, however, not a trivial task.[45] Given the explosion of computational cost with the size of the system for the more accurate post-Hartree–Fock methods, most first-principles calculations in heterogeneous catalysis are done with DFT. The key challenge within the DFT formalism is the exchange–correlation functional since the exact form is still unknown. Hence, different approximations made to define such a functional lead to a hierarchy of DFT functionals with increasing complexity and computational requirements. Due to the computational needs imposed by system sizes, many of the calculations in heterogeneous catalysis are performed with semi-local functionals. They often yield a decent account of covalent bonds and geometric structures but they face some issues. One of them is the spurious electron delocalization that arises from an incomplete cancellation of repulsive Coulomb self-interaction contributions by the approximate exchange energy given by the employed functional. This artificial delocalization promotes, for instance, overestimation of the binding of gas-phase molecules relevant to catalysis (*e.g.*, the PBE functional can overestimate O_2 binding energy by 1 eV (ref. 46)). Thus, it is important to be cautious when modelling the interaction of such molecules with solid surfaces since it is primarily error cancellation that allows a reasonable accuracy (~0.1–0.3 eV) to be reached. Another challenge comes from the pseudopotential (PP) used to model the electrons.[47] PPs are used in plain-wave (PW)-based calculations for smearing the nuclear charge and modelling the core electrons. Each PP contains the minimum energy cutoff that might be used in the calculations, characterized by a "core radius". Depending on the magnitude of such a core radius, there are "hard" (with a small core radius, needing more PW basis functions and higher cutoff energies for describing the region beyond the core radius) or "soft" (with a larger core radius, requiring lower cutoff energies and fewer PW basis functions) PPs. Although soft PPs are less computationally demanding, too large a core-radius could compromise the quality of the calculations. However, it has been demonstrated that the use of ultrasoft PPs or those based on the projector augmented-wave (PAW) method are in good agreement with those from all-electron calculations.[48]

Regardless of the level of theory employed to perform calculations, the value of the E_{ads} depends on the site where the atom/molecule gets adsorbed during the reaction. Different surfaces have different adsorption sites (for instance, see Fig. 2 in ref. 49). The most probable site where an adsorbate will be adsorbed is the one that is more energetically favorable (*i.e.*, the one featuring the most negative E_{ads}). Hence, the characterization of the E_{ads} of an adsorbate can be done, in the most general manner, based on the identity of the catalyst and the adsorption site. In a previous study,[49] we defined a set of geometric and electronic features that describe different properties associated with the identity of the catalyst and the adsorption site: the generalized coordination number (GCN),[50] Ψ,[51] the weighted atomic radius (WAR), weighted electronegativity (WEN), weighted first ionization energy (WIE), and the outer electrons.

The GCN and Ψ describe the geometry configuration of the adsorption site and the chemical environment around the site, respectively, and can be obtained as

$$GCN = \frac{\sum_{i=1}^{n} CN_i}{CN_{max}} \quad (4)$$

where CN_i is the coordination number of the i-th first or second nearest neighbor of the adsorbate, N is the total number of atoms composing the sets of first and second neighbors, and CN_{max} is the maximum coordination number for a given crystal lattice;

$$\Psi = \frac{\left(\prod_{i=1}^{n} S_i\right)^{\frac{2}{N}}}{\left(\prod_{i=1}^{n} EN_i\right)^{\frac{1}{N}}} \quad (5)$$

where N is the number of atoms at active centers, whereas S_i and EN_i are the outer electrons and the Pauling electronegativity of the i-th atom at active centers.

From the remaining features, WEN, WIE, WAR, and the outer electrons are related to the identity of the material and how the stoichiometry affects relevant electronic and geometric properties.

We used these features to construct a dataset of the E_{ads} of H, O, and OH on several binary intermetallic materials and pure metals calculated with DFT calculations for the training of ML models (Fig. S1 in the SI histograms of E_{ads} for each of the datasets where it is evident that they provide a good representation of different ranges of the target property). In addition to the aforementioned features, we also included the unit cell volume of the bulk crystal and the strain (we computed the E_{ads} considering the application of different elastic strains on the surfaces). Although additional descriptors (such as the d-band center and the d-band width) can provide a better description of the local chemical environment,[52–54] they require expensive DFT calculations that would hinder an efficient utilization of any trained ML model.

This dataset was used in this work to train the ML models used to test our XAI strategy for materials discovery and design. Further details of the dataset can be found in Section S1 in the SI and the original publication.[49]

2.2 XAI materials design and the discovery strategy

We consider that for an XAI design and discovery strategy to be successful it needs to fulfill certain requirements: (i) be founded on an ML model that provides accurate predictions of the target property, (ii) such predictions need to be reliable (*i.e.*, are obtained through rigorous chemical and physical grounds and are not the result of spurious correlations), and (iii) the design and discovery task needs to provide explanations on why a molecule/material is a good candidate compared to others. Because of this, we propose the workflow shown in Fig. 1a. It consists of three main steps, model selection, model validation, and the XAI design, each accounting for one of the aforementioned requirements. In the following, we are going to briefly summarize the specific tasks we performed in each of the parts of the workflow. However, please note that our workflow is very flexible and that the methods and tools employed within each block (green boxes in Fig. 1a) can be changed by others.

2.2.1 Model selection. The first part of any ML application is the construction of the model. The objective here is to find the most accurate model for predicting the target property. For achieving this, one needs to test different methods trained on an equal footing (*i.e.*, same dataset, same training scheme, assessed with the same metrics, *etc.*).

In the context of this work, we used as our test example the search for heterogeneous catalysts for hydrogen production and energy generation. According to the Evans–Polanyi[55] and the Sabatier principle,[42] the adsorption energies (E_{ads}) of the intermediate species determine the suitability of a material to catalyze a given reaction. As mentioned before, the reference dataset that we used is the one we recently published containing DFT E_{ads} of H, O, and OH on binary intermetallic compounds (A_xB_y) and pure metals (a description of the dataset can be found in Section S1 in the SI).[49] The E_{ads} is a continuous variable and, hence, we explored the performance of different regression models in the prediction of this property. Namely, we tested extremely randomized trees regression (ET), XGBoost, Gaussian process regression (GPR), kernel ridge regression (KRR), and feed-forward neural networks (NNs). The resulting models were compared through the typical mean absolute error (MAE) and root-mean squared error (RMSE) metrics, and their performance was visualized with parity plots. The best performing model is the one used in the next steps of the workflow.

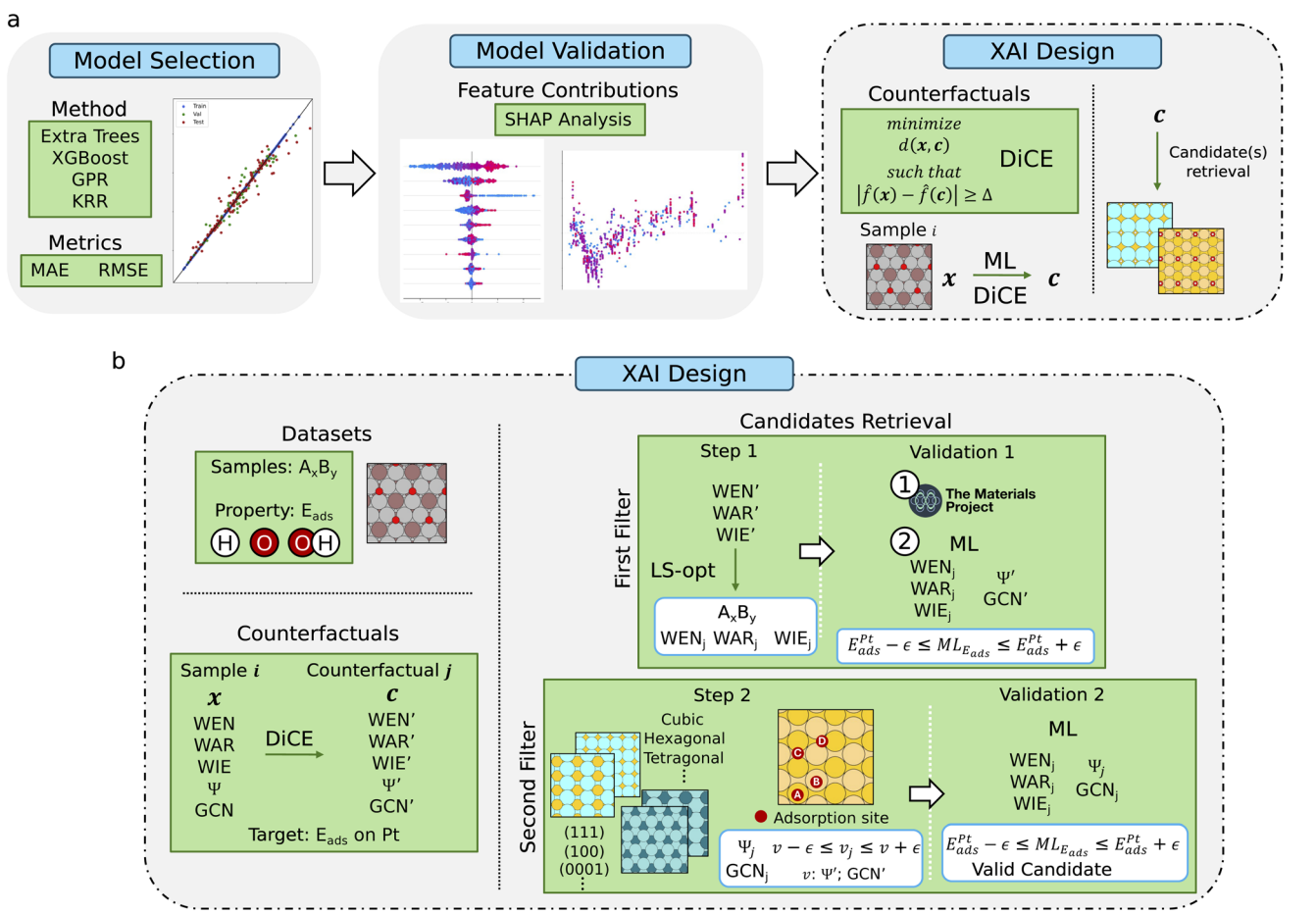

Fig. 1 (a) General workflow of the proposed XAI discovery and design strategy. The workflow is flexible so that the methods and tools employed within each block (green boxes) can be changed easily by others. (b) Procedure followed for the retrieval of candidates within the XAI design step. After counterfactuals are generated for each of the samples available in the dataset, the candidates are recovered from the set of new features. Explainability is ensured by construction since every candidate found can be linked to the original sample.

2.2.2 Model validation. Here we do not employ the term "validation" to refer to the process of assessing performance on unseen data to fine-tune the ML model. Instead, in this step of the workflow, we intend to make a validation of the model by chemical and/or physical means. The objective is to have more certainty that the prediction accuracy reached by the selected model in the previous step is not the result of spurious correlations, but the consequence of the model learning chemical/physical principles. To this end, we propose the use of feature contributions for explaining how the ML model obtains its predictions. Specifically, in this work, we chose to analyze feature contributions by Shapley values[56] using SHapley Additive exPlanations (SHAP).[57] This is a method of computing feature importance weights as a complete explanation (a broader description of Shapley values can be found in Section S1 in the SI).

We checked which were the features that contributed the most to the predictions of our models and analyzed their chemical/physical significance. For instance, one can expect that features such as the area of the adsorption site and the electronegativity might play an important role in the adsorption processes involved in catalysis. Hence, we would expect that a ML model can capture such correlations whenever the appropriate information is fed into the descriptor. Once the model is validated through the analysis of feature contributions, the model is deployed for the design and discovery step.

2.2.3 XAI design. Our proposed XAI design strategy is based on the use of counterfactual explanations. These explanations are a tool that provides insights into what changes in input features would lead to a different prediction outcome. In their most simple form, counterfactuals are defined, for a regression task as ours, using the solution of the following constrained optimization problem:[58]

$$\begin{aligned} &\text{minimize } \mathrm{dist}(\mathbf{x},\mathbf{c}) \\ &\text{such that} \left|\hat{f}(\mathbf{x}) - \hat{f}(\mathbf{c})\right| \geq \Delta \end{aligned} \tag{6}$$

where $\mathbf{x}$ is the feature vector of a known sample of the dataset, $\mathbf{c}$ is the feature vector of the counterfactual, dist($\mathbf{x}$,$\mathbf{c}$) is a measure of the distance between features and $\hat{f}$ is the ML model, while Δ is a problem specific hyperparameter which denotes the desired change in the value.

In this work, we employ Diverse Counterfactual Explanations (DiCE)[59] for the generation of counterfactuals. DiCE not only satisfies the feasibility/proximity property (minimization of dist($\mathbf{x}$,$\mathbf{c}$)), but also the diversity of counterfactuals to provide different ways of changing the prediction outcome. This is achieved by optimizing a combined loss function over all generated counterfactuals:

$$\begin{aligned} C(\mathbf{x}) = \underset{\mathbf{c}_1,\ldots,\mathbf{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} y\mathrm{loss}(f(\mathbf{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k}\mathrm{dist}(\mathbf{c}_i,\mathbf{x}) \\ - \lambda_2\ \mathrm{dpp\_diversity}(\mathbf{c}_1,\ldots,\mathbf{c}_k) \end{aligned} \tag{7}$$

where $\mathbf{c}_i$ is a counterfactual example, k is the total number of counterfactuals to be generated, f(.) is the ML model, $y$loss(.) is a metric that minimizes the distance between the prediction of f(.) for $\mathbf{c}_i$ and the desired outcome y, $\mathbf{x}$ is the original input, dist(.) denotes a distance metric between the counterfactual and the original input, dpp_diversity(.) is the diversity metric, and λ_1 and λ_2 are hyperparameters that balance the three parts of the loss function (we refer the readers to the original publication of DiCE in ref. 59 for a more detailed description).

As a result of the optimization problem, a new set of features is available for each k-th $\mathbf{c}$, whose prediction outcome is equal, or at least, close to the defined target value. However, DiCE cannot find the identity of the sample. Hence, the next step in our XAI design strategy is to transform back each $\mathbf{c}$ from feature space to chemical compound space. The procedure to do it is dependent on the space covered by the dataset and on the features used to generate the counterfactuals. Therefore, it will change for different datasets and applications. Please note that this peculiarity is also found in other design approaches. In generative models, for instance, the latent space in which the models seek for new samples is bound to the chemical space learned by the model (*i.e.*, the space covered by the dataset) and the process for retrieving the identity of such samples will always change depending on the specific features in the dataset.

Here, we followed the procedure summarized in Fig. 1b for retrieving new candidates from the generated counterfactuals. We carried it out for each adsorbate separately to have three different tests for our XAI design strategy. The features of the dataset that we used to generate the counterfactuals can be gathered in two groups: composition- (weighted electronegativity [WEN], weighted atomic radius [WAR], and weighted ionization energy [WIE]) and adsorption-related (generalized coordination number [GCN] and Ψ) features. From the former we recovered the stoichiometry of the counterfactual candidate, whereas the latter were used to obtain the information about the crystal lattice, the surface facet, and the adsorption site. Although the adsorption site is not part of the identity of the candidate, it is needed for validating that the E_{ads} is within the design target. We selected as the target for each adsorbate the lowest E_{ads} on a Pt fcc (111) surface as found in the reference dataset ($E_{\mathrm{ads}}^{\mathrm{Pt}}$). We considered such a target suitable for our tests because Pt-based catalysts are the best performing for the production of hydrogen and energy generation. Specifically, we set the range of the target as $E_{\mathrm{ads}}^{\mathrm{Pt}} \pm 0.15$ eV for O and OH and $E_{\mathrm{ads}}^{\mathrm{Pt}} \pm 0.10$ eV for H. We did not use the number of valence electrons of the atoms involved in the material (also available in the dataset) for counterfactual generation because, when used, DiCE converged most of the time to pure Pt or Pd.

Once we generated the counterfactuals, the retrieval of candidates was done in two steps, each of which was followed by a validation stage. In the first step of the retrieval process, we performed least-squares (LS) optimization on the following set of equations to find the stoichiometry of the material (*i.e.*, subscripts x and y in A_xB_y):

$$\begin{aligned} &\mathrm{EN_A}X + \mathrm{EN_B}Y = \mathrm{WEN}' \\ &\mathrm{AR_A}X + \mathrm{AR_B}Y = \mathrm{WAR}' \\ &\mathrm{IE_A}X + \mathrm{IE_B}Y = \mathrm{WIE}' \end{aligned} \tag{8}$$

where EN is the electronegativity, AR is the atomic radius and IE is the ionization energy, while X and Y are the proportions of A and B in the candidate, respectively, and the values for which we

solved the LS optimization. Therefore, the values of x and y were equal to the smallest integers holding the proportions (*e.g.*, if $X = 0.34$ and $Y = 0.66$, then $x = 1$ and $y = 2$). WEN′, WAR′, and WIE′ are the weighted electronegativity, the weighted atomic radius, and the weighted ionization energies from the counterfactual. As we kept constant the number of electrons of the original sample when generating the counterfactuals, we solved eqn (8) considering all possible pairs of A and B. For instance, if the number of valence electrons in the original sample was 12 for A and 4 for B, we found the solution of eqn (8) for all valid (*i.e.*, atoms that were available in at least one material in the dataset) pairs between groups 12 and 4 of the periodic table: Zn_xTi_y, Zn_xZr_y, Zn_xHf_y, Cd_xTi_y, Cd_xZr_y, and Cd_xHf_y. After obtaining a stoichiometry, we validated the candidate through two filters. The first one consisted in confirming the existence of the candidate by searching in the Materials Project database.[60] If the candidate existed, we updated the WEN, WAR, and WIE values of the counterfactual with the real values from the candidate. The stoichiometry was obtained through LS optimization and, hence, the features are not exactly the same as the ones found during counterfactual generation. Because of this, as a second filter we used our ML model to predict the E_{ads} of the updated counterfactual. If the predicted E_{ads} remained within the $E_{ads}^{Pt} \pm \varepsilon$ range (where ε is a predefined threshold. In our case, ε was set to 0.10 eV for H and 0.15 eV for O and OH), the candidate was kept for the second step of the retrieval of candidates.

For all valid candidates obtained in the first step, we considered all possible crystal lattices (cubic, hexagonal, tetragonal, *etc.*) as found in the Materials Project and constructed surface slabs for different ($hk(i)l$) facets (*e.g.*, (111), (0001), and (100)) using the Atomic Simulation Environment (ASE).[61] We tried to keep everything as simple as possible, so we only considered the most common (*i.e.*, small numbering) facets. We set the highest integer for h, k, and l equal to 2 for hexagonal lattices (knowing that $i = -(h + k)$) and 3 for all other lattices. The surfaces were constructed with 4 layers as most surfaces in the reference dataset. We discarded all surfaces that had large gaps between atoms and those with a unit cell containing more than 32 atoms (8 atoms per layer). Then, we identified all available adsorption sites on the selected surfaces and computed for each of them the values of the GCN (eqn (4)) and Ψ (eqn (5)).

Finally, we compared the GCN and Ψ values of each adsorption site to the GCN and Ψ values of the counterfactual. If both values lied within a predefined threshold from the counterfactual values (we set these thresholds to be 20% of the counterfactual value for the GCN and 15 units for Ψ), we updated the values of the GCN and Ψ of the counterfactual with the values from the actual candidate. For validating the final candidates, we used again our ML model to predict the E_{ads} for the updated set of features. Analogously to the first validation stage, we only kept the information (stoichiometry, crystal lattice, facet, and adsorption site) of those candidates whose predicted E_{ads} was within the $E_{ads}^{Pt} \pm \varepsilon$ range (in this second validation, we kept ε equal to 0.10 eV for H, but we were more stringent with O and OH by also setting ε to a value of 0.10 eV). The set of candidates obtained after this validation stage was the outcome of our XAI design strategy. DFT calculations were then performed to confirm their E_{ads}.

2.3 ML models for the prediction of adsorption energies

We tested several ML methods (ET, XGBoost, GPR, KRR, and NNs) in order to obtain the most accurate models for predicting the E_{ads} of H, O, and OH from the dataset of binary intermetallic compounds (A_xB_y) and pure metals. We constructed one model for each adsorbate separately to have three different design tests for our XAI design strategy. All our models were constructed using Scikit-learn,[62] except for the XGBoost method that was used as implemented in the xgboost Python package[63] and the NNs, that were constructed with Keras.[64] We split our dataset with a "pseudo-random" procedure into 85% train/validation (783 samples for H, 757 for O, and 687 for OH) and 15% test (138 samples for H, 134 for O, and 121 for OH). The train/validation partition was then used for hyperparameter optimization (the list of optimal hyperparameters and the grids used for optimization can be found in Section S2 in the SI) through a "pseudo-randomized" grid search 10-fold cross-validation, except for the NNs, for which we used the Bayesian optimization search as implemented in KerasTuner.[65] We use the term "pseudo-random" here because the different sets were not constructed fully randomly. Similar to what was done in ref. 49, we decided to have in our training sets all samples of pure metals and all samples of single-element adsorption sites (*e.g.*, ontop-A and longbridge-B) for which there were no pure surfaces of the given element. For instance, there is no pure Al in the dataset and we forced all ontop-Al samples into our training sets. We justify this approach because the E_{ads} on pure metal surfaces are already known and on the single-element positions of the metals without pure surfaces the adsorption is normally very energetically unfavorable. However, they do provide valuable information about the adsorption processes for our ML model. Please note that after adding these initial samples to the training sets, the remaining samples for training, as well as the samples for validation and testing, were randomly selected.

Table 1 shows the average mean absolute errors (MAEs) and root-mean squared errors (RMSEs) on the test sets over all cross-validation tasks of all the tested methods with optimized

Table 1 Average E_{ads} prediction accuracy on the test set for all methods used in this work over all cross-validation tasks in terms of MAE and RMSE for H, O, and OH. Errors are in eV. Bold font is used to indicate the lowest error among the methods

Adsorbate	Metric	ET	XGBoost	KRR	GPR	NN
H	MAE	**0.10**	0.11	0.13	0.15	0.18
	RMSE	**0.21**	**0.21**	0.24	0.25	0.26
O	MAE	**0.25**	0.27	0.34	0.34	0.33
	RMSE	**0.43**	0.47	0.54	0.55	0.52
OH	MAE	**0.20**	0.23	0.27	0.41	0.33
	RMSE	**0.33**	0.37	0.42	0.60	0.49

hyperparameters (the values of the hyperparameters are available in Section S2 in the SI). For the case of H, all tested methods performed similarly, with differences of only 0.08 eV in MAE and 0.05 eV in RMSE between the best (ET) and the worst (NN) methods. Indeed, both the ET (MAE = 0.10 eV; RMSE = 0.21 eV) and the XGBoost (MAE = 0.11 eV; RMSE = 0.21 eV) methods have practically the same prediction accuracy. The trend observed in terms of accuracy is ET < XGBoost < KRR < GPR < NNs. Such a trend slightly changes for the prediction of the E_{ads} of O, since the NN is now more accurate than the KRR and GPR models. KRR and GPR present the worst performance with a MAE as high as 0.34 eV and RMSEs of 0.54 and 0.55 eV, respectively. The ET was again the best performing method with a MAE and RMSE of 0.25 and 0.43 eV, respectively. Finally, for OH, ET was still the method with the lowest prediction errors (MAE = 0.20 eV; RMSE = 0.33 eV). GPR was the worst method as with the other adsorbates with errors that doubled those of ET, while KRR presented a higher accuracy than the NN model. Henceforward, we will only focus on the ET method, which was the most accurate for the prediction of E_{ads} for all the adsorbates. Please note that we later explored a more refined grid than that shown in Section S2 in the SI for the number of fitted sub-estimators in the ET method to ensure that we had indeed the best possible model. We found that there is a saturation behavior, where performance gains seem to diminish rapidly after a certain point and no models with a better performance than those in Table 1 were obtained.

Although the prediction accuracy of the ET method for the three adsorbates is good, we decided to take a look in more detail. For this, we show in Fig. 2 parity plots of the best ET model out of all cross-validation tasks for each adsorbate. The performance of the best models shows that MAEs are very close to the average ones presented in Table 1, while RMSEs are slightly better (up to 0.03 eV for the prediction of OH E_{ads}). This means that the prediction accuracy of an ET model trained on our reference data is quite consistent regardless of the specific training set employed for its construction. Also, all tested samples are uniformly scattered to the sides of the perfect prediction baseline, as confirmed with the R^2 values. For the prediction of the E_{ads} of H, the R^2 is low compared to those obtained for O and OH. However, this is the consequence of the scale of E_{ads} for H, which is smaller than those of O and OH.

Furthermore, since it is possible to have models with good overall prediction metrics that are considerably less accurate on certain samples (*i.e.*, outliers), we were also interested in ensuring that such a problem was minimal in the case of our ML models. For this, we defined that an outlier is a sample for which the AE in prediction is larger than the MAE + 3σ, where σ is the standard deviation of all AEs on the test set (σ was equal to 0.16, 0.33, and 0.23 eV for H, O, and OH, respectively). As can be seen in Fig. 2, the three models perform really well with only a few test (3, 4, and 4 for H, O, and OH, respectively) and validation (2, 0, and 3 for H, O, and OH, respectively) outliers.

2.4 Validation of ML models through chemical/physical concepts

We have shown that the ET method was the best performing among the four methods explored, providing reliable predictions of the E_{ads} for all adsorbates. Nevertheless, we wanted to be sure that those predictions were obtained by leveraging chemical/physical concepts and not by the exploitation of spurious correlations. To this end, we employed SHAP[57] to compute feature contributions and, hence, explain how the ML models obtained their predictions. The ML models used for this

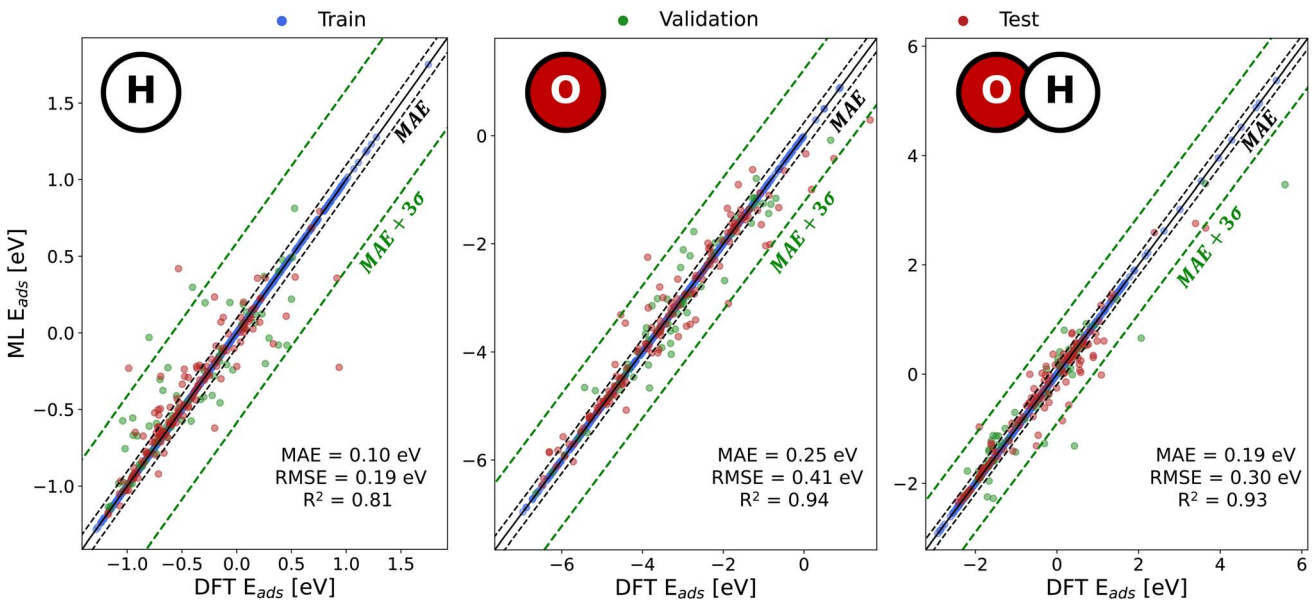

Fig. 2 Parity plots of the best ET models out of all cross-validation tasks for the prediction of the E_{ads} of H, O, and OH. Black dashed lines delimit the region of errors equal to or lower than the MAE and the green dashed lines delimit the region where errors are equal to or lower than the MAE plus 3 times the standard deviation (σ) of AEs on the test set.

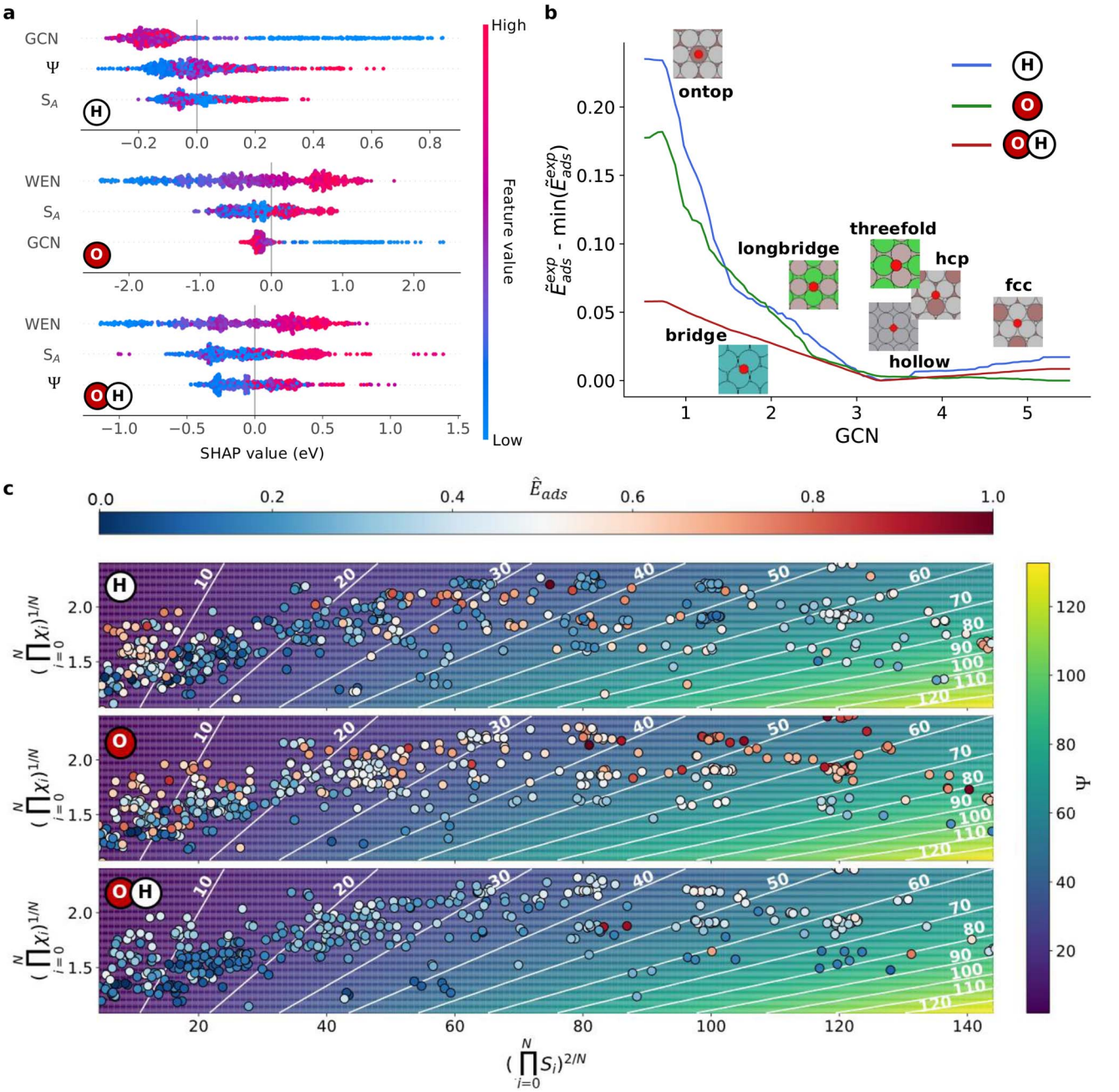

Fig. 3 Analysis of the most relevant features. (a) Beeswarm plots of SHAP values for the three most important features for each adsorbate as obtained from SHAP analysis; (b) partial dependence plots of E_{ads}, in terms of the relative normalized expected E_{ads} ($\tilde{E}_{ads}^{exp} - \min(\tilde{E}_{ads}^{exp})$), on the GCN for each adsorbate; and (c) Ψ value heatmaps for each adsorbate, where all unstrained samples in the datasets are plotted with the components of Ψ (eqn (5)) and colored according to their normalized E_{ads} ($\hat{E}_{ads}$). A small random displacement is applied to all points for the sake of clarity.

task were ET models (one for each adsorbate) trained on all samples of our datasets using the set of optimal hyperparameters (see Table S1 in the SI) of the models discussed in the previous section. Training on all dataset samples is justified for model deployment. For the XAI discovery and design, we require the best possible model to ensure accurate predictions of the unknown set of features generated by DiCE. For generating the SHAP explainer model we employed 15% of the dataset. The samples were selected using the internal routine for sampling available in the SHAP package.

Fig. 3a shows beeswarm plots of the three most important features for predicting the E_{ads} of each adsorbate according to the SHAP analysis (plots containing all other features of the dataset are in Fig. S2–S4 in the SI). If the value of a feature is related to a positive SHAP value, it means that such value leads to an increment of the prediction outcome (*i.e.*, a more positive E_{ads} in our case), while a negative SHAP value indicates a reduction of the prediction outcome (*i.e.*, a more negative E_{ads}). We observe that the sets of features are similar between the three adsorbates. In fact, the number of valence electrons in

atom A (S_A) is relevant for the three adsorbates (the second most important feature for both O and OH, and the third one for H). The SHAP analysis indicates that, in general, the larger the number of electrons the more positive the E_{ads} is. We can relate this to the downshifting of the d-band center with an increasing number of valence electrons.[66] It is well known that a more negative d-band center leads to more positive E_{ads},[67] making adsorption less favorable for the three adsorbates. The other important features, WEN, GCN, and Ψ, are each shared between two adsorbates.

In the case of WEN, it is the most important feature for the ML model to predict the E_{ads} of O and OH. The relevance of this feature is supported by general chemistry knowledge. O is one of the most electronegative species (3.44 on the Pauling scale). Hence, we can expect it to bind stronger to atoms with lower electronegativity. This is exactly what can be concluded from our SHAP analysis. The SHAP values of surfaces with a low WEN are very negative (as low as below −2.0 and −1.0 eV for O and OH, respectively) and very positive (up to around 1.0 eV for both adsorbates) for samples with a high WEN.

As per the GCN, which is the most important feature for the prediction of the E_{ads} of H and the third one for O, we also see a clear trend. Whenever the GCN is small the adsorption becomes quite unfavorable (see the positive SHAP values for both H and O), while it is slightly more favorable for large values of the GCN. This can be explained by looking at the definition of the GCN in eqn (4). The value of this feature is dependent on the coordination of the nearest neighbors of the adsorbate on the surface. The number of such neighbors and their position on the surface vary across the different available adsorption sites. Therefore, the GCN implicitly indicates the type of site where the adsorbate is adsorbed. Lower values of the GCN represent sites with lower coordination (*e.g.*, a GCN of around 1 is linked to ontop positions, whereas values from 1.5 to 2.0 and from 2.0 to around 3.0 correspond to (short)bridge and longbridge sites, respectively). On the other hand, highly coordinated adsorption sites, such as threefold, hcp, hollow, and fcc show GCN values above 3.5, with fcc sites being those with the highest GCNs (around 5.0). To analyze this with further details, Fig. 3b shows partial dependence plots on the GCN for the three adsorbates. We represent the impact on E_{ads} as the relative normalized expected E_{ads} ($\tilde{E}^{exp}_{ads} - \min(\tilde{E}^{exp}_{ads})$), where E^{exp}_{ads} corresponds to the average model prediction given a value of the GCN. For calculating it, we first applied a min–max normalization to the E^{exp}_{ads} values for each adsorbate to obtain a normalized E^{exp}_{ads} ($\tilde{E}^{exp}_{ads}$). The minimum and maximum values for normalization employed for each adsorbate were those of the minimum and maximum E_{ads} in their corresponding datasets (more details on the procedure can be found in Section S1 in the SI). Subsequently, a minimum-value (*i.e.*, $\min(\tilde{E}^{exp}_{ads})$) shift was applied to each value of $\tilde{E}^{exp}_{ads}$, separately. Fig. S7, in the SI, shows a graph of the data in Fig. 3b before processing to understand the applied transformation and the need to do it. The use of a relative and not the actual value of $\tilde{E}^{exp}_{ads}$ is to compare the behavior of model predictions for all the adsorbates from the same baseline. As suggested by the SHAP values (Fig. 3a), low GCN values lead to considerably higher E_{ads} (as great as 20% of the range length of E_{ads}) than those predicted for high GCNs for H and O. This means that the ML model has correctly learned that ontop positions are normally the least favorable ones, while fcc (for O) and threefold, hcp, and hollow (for H) sites are usually the most energetically favorable. Furthermore, our partial dependence plot allows us to understand why the GCN is not as relevant for OH as it is for the other adsorbates. The differences in $\tilde{E}^{exp}_{ads}$ are just 0.05 between the sites with low and high coordination.

Finally, the Ψ feature is found to be of high importance for the prediction of the E_{ads} of H and OH. The impact on model output is the contrary to that of the GCN. Low values of Ψ usually correspond to more negative E_{ads}, whereas high values of the feature are linked to a less favorable adsorption. The interpretation of Ψ (see eqn (5)), however, is more complex than that of other features. Ψ can be considered a description of the chemical environment of the specific binding position of the adsorbate by means of the electronic properties of the nearest neighbors. Specifically, the chemical environment is described using the geometric means of the electronegativity $\left(\left(\prod_{i=1}^{N} EN_i\right)^{\frac{1}{N}}\right)$ and the number of electrons $\left(\left(\prod_{i=1}^{N} S_i\right)^{\frac{2}{N}}\right)$ of the surface atoms at the adsorption site. We plot in Fig. 3c all samples in our dataset with the strain descriptor equal to zero (*i.e.*, unstrained samples) as a function of the aforementioned variables on a heatmap of Ψ values for the three adsorbates. Each point in our plots is colored according to a normalized E_{ads} ($\tilde{E}_{ads}$). From the plots, it can be concluded that, for H and OH, adsorption sites with Ψ values between 10 and 30 tend to have, in general, lower E_{ads} than sites with larger or smaller Ψ values. In particular, a range between 10 and 20 seems to be the most favorable for adsorption. Apart from the large E_{ads} that are found for Ψ values below 10, the results observed are in agreement with the trends provided by the SHAP analysis (Fig. 3a). For O adsorption, our plots allow us to explain why Ψ is not among the three most important features for the prediction of E_{ads}. It is evident that the plot could be divided into two sections by a geometric mean electronegativity of 1.75. Above this value, the E_{ads} of O are quite larger than those below. This indicates that, regardless of the value of Ψ of the adsorption site, the electronegativity of the species on the surface will be the one defining how strongly the O atom will bind to the surface. A similar trend can be found for OH for values of Ψ beyond 60 (in agreement with WEN being the most important feature for describing OH adsorption and Ψ the third one), while for H such a trend is absent (*e.g.*, for a Ψ of around 50, we observe both high and low E_{ads} for different electronegativities). It is important to remark that some of the general trends of E_{ads} with respect to changes in Ψ discussed here were found previously.[51]

2.5 XAI for designing and discovering new materials

After confirming that our ML models provided accurate, reliable, and rigorous predictions, we used them for testing our strategy for the designing and discovery of new materials through XAI. As mentioned before, we propose the use of

Table 2 Candidates discovered and designed with the proposed XAI strategy. Both the ML predicted and reference DFT E_{ads} are shown. Bold font is used for those candidates that were confirmed as valid after the DFT reference calculations (*i.e.*, the difference between ML E_{ads} and DFT E_{ads} was below the MAEs shown in Fig. 2). "—" indicates that the adsorption site was found to be unstable. The designing targets were −0.49, −1.79, and 1.19 eV for H, O, and OH, respectively. The ML E_{ads} shown here is the one obtained using our ML model on the final candidate (*i.e.*, after all filters are passed)

Adsorbate	Material	Crystal	Facet	Site	ML E_{ads} (eV)	DFT E_{ads} (eV)
H	CaNi	Cubic	101	Bridge-CaNi	−0.46	—
	YIr	Cubic	101	Ontop-Ir	−0.45	−0.68
	YAu	**Cubic**	**100**	**Hollow**	**−0.47**	**−0.42**
	YAu	Cubic	101	Hollow	−0.53	—
O	**ZnIr**	**Hexagonal**	**10−11**	**Hollow**	**−1.72**	**−1.93**
	ZnPt	**Tetragonal**	**101**	**Hollow**[a]	**−1.76**	**−1.81**
OH	**CdAu**	**Cubic**	**101**	**Longbridge-Au**	**1.10**	**1.06**
	CdAu	Hexagonal	0001	Bridge-Au	1.10	0.89

[a] After geometry relaxation the ZnPt bct(101) surface suffered from modifications. We used the same coordinates found for the hollow site in the candidates' retrieval process (based on the unrelaxed surface) to place the O atom on the relaxed slab. After geometry relaxation, the adsorbate relaxed on a bridge-Pt site (see Fig. S8 in the SI for the images of the original and relaxed surface and O site).

counterfactual explanations as an XAI tool to find new candidates with desired properties and, at the same time, retrieve explanations for helping in understanding why certain materials are better than others. Specifically, we employed DiCE[59] to generate the counterfactual explanations. As inputs for DiCE, we used all unstrained samples in our dataset and set the maximum number of generated counterfactuals to 4 for each sample. We selected this number to achieve a trade-off between computational cost and the amount of generated candidates. The target in our discovery and design task was the E_{ads} of each adsorbate on the most energetically favorable site on a Pt fcc (111) surface as found in the dataset (−0.49 eV for H, −1.79 eV for O, and 1.19 eV for OH). Among the available approaches in DiCE for counterfactual generation, we selected the genetic algorithm method for the cases of H and O, and, due to computational efficiency, the randomized sampling for OH.

In total, we were expecting to generate more than 1600 counterfactuals for each adsorbate. From this pool of counterfactual explanations and after applying the validation filters summarized in Fig. 1b, we found 8 candidates (*i.e.*, material + facet) featuring the desired target E_{ads}: 4 for H, 2 for OH, and 2 for O. Even though our search for candidates was successful using the default weight hyperparameters of DiCE (λ_1 and λ_2 in eqn (7)), we explored the use of different values from the default ones (0.5 and 1, respectively) on the example of H. However, we found that, for our case, the generated counterfactuals did not vary significantly between different values of λ_1 and λ_2 and no additional candidates were identified. In any case, we suggest that for different applications, the effect of λ_1 and λ_2 in the generation of counterfactuals should be assessed. All the candidates found were then validated with DFT calculations (the computational details of the calculations are given in Section S1 in the SI). Table 2 shows all relevant information about the discovered candidates through our XAI strategy, as well as the ML predicted and the DFT E_{ads}. We considered a candidate to be valid if the difference between the predicted E_{ads} and the DFT E_{ads} was within the MAE shown in Fig. 2 (0.10, 0.25, and 0.19 eV for H, O and OH, respectively). Out of the 8 candidates, 4 of them were confirmed using DFT calculations: YAu bcc(100) for H, ZnIr hcp(10−11) and ZnPt bct(101) for O, and CdAu bcc(110) for OH. Considering the designing targets (−0.49, −1.79, and 1.19 eV for H, O, and OH, respectively), the largest error among the confirmed candidates was of only 0.14 eV for ZnIr hcp(10−11), hence demonstrating the potential of our proposed approach. Moreover, it is important to highlight that we did not have any sample in the dataset for the bcc(100) and hcp(10−11) facets and for the tetragonal crystal lattice. We proved before (see Fig. 3) that our ML models learned sensible chemical/physical concepts from our data, and their applicability to unknown crystal structures and surfaces further confirms this. This knowledge is then successfully exploited within our XAI approach.

As for the candidates that were not confirmed through DFT calculations, there were two of them, YIr bcc(110) for H and CdAu hcp(0001) for OH, for which the differences between DFT and ML E_{ads} were higher than the expected accuracy of the models. In both cases, the ML model underestimated the binding between the adsorbate and the surface at the discovered site by more than 0.20 eV. In the case of YIr bcc(110), the site found for this candidate was ontop-Ir and, following the trends shown in Fig. 3b; the model probably assumed that the low GCN of the site would lead to a more unfavorable environment for the H atom than it actually is. For CdAu hcp(0001), the deviation between the reference DFT E_{ads} and the predicted one can be related to the fact that (short)bridge positions for OH are usually unstable (*i.e.*, they are not a minimum in the potential-energy surface) in binary intermetallic surfaces (as was found during the construction of the dataset[49]) leading to an under-representation of these sites in the dataset. Still, the error (0.21 eV) is lower than the RMSE (0.30 eV) of the models assessed in Fig. 2. Given the finding of these large errors between the DFT E_{ads} and the ML E_{ads}, we explored the possibility of using the variance of the predictions of the individual sub-estimators of our ET models as an additional filter prior to DFT calculations. We took as examples the candidates in Table 2 from which the identified adsorption sites were found to be

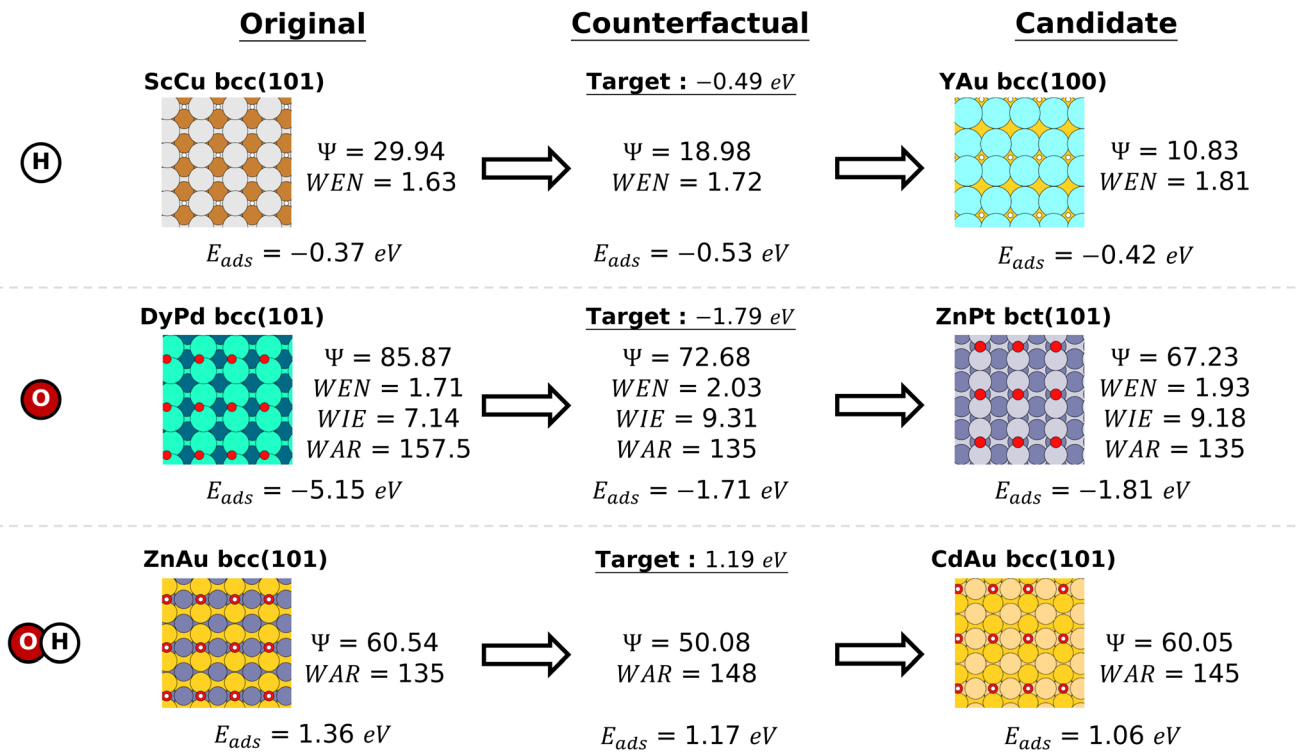

Fig. 4 Change of feature values from original samples to counterfactuals to validated candidates. The data are shown for three of the discovered surfaces: YAu bcc(100) for H, ZnPt bct(101) for O, and CdAu bcc(101) for OH. The E_{ads} shown for the original samples and the candidates are the E_{ads} calculated with the reference DFT calculations, while the one of the counterfactuals is an E_{ads} predicted with our ML model for the original counterfactual (*i.e.*, before going to the candidates' retrieval step). The units of the WAR are pm and those of WIE are eV.

stable (YIr bcc(110) and YAu bcc(100) for H, ZnIr hcp(10−11) and ZnPt bct(101) for O, and CdAu bcc(101) and CdAu hcp(0001) for OH). We found (see Table S6 in the SI) that, for H and O, there is indeed a correlation between the variance of the fitted sub-estimators and the error of the ML model with respect to DFT but it is not trivial to define what is a large or small variance, while for OH, both CdAu candidates show the same variance despite the rather large difference in accuracy of the ML model. This small exercise shows that the variance of sub-estimators in ensemble ML models could be used as an additional filter to select potential candidates. Nevertheless, its applicability is conditioned on the definition of a proper threshold for each adsorbate. This requires further research and a deeper analysis into the performance of ML models on a larger number of samples. Although it is worth exploring this, it is out of the scope of the present work.

For the remaining two candidates that were not confirmed, CaNi bcc(110) and YAu bcc(110), the adsorption sites found (bridge-CaNi and hollow, respectively) resulted in being unstable (*i.e.*, the adsorption site changed during geometry relaxation). This is not a failure either of the proposed XAI strategy or of the ML models used. The ML models are only constructed for the task of predicting E_{ads} and, hence, cannot foresee the instability of an adsorption site. Analogously, the XAI design and discovery strategy rely on the available ML models only and do not include any DFT- or ML-based filter to help it anticipate possible unstable adsorption sites. The implementation of such a filter in the pipeline of our proposed strategy is possible but we considered that it is out of the scope of the present work.

We have demonstrated the potential of our XAI strategy for the design and discovery of new materials but nothing has been discussed yet regarding the explanations that we can recover from it. In this regard, Fig. 4 shows the features of the original dataset samples, the generated counterfactuals, and examples of the discovered and validated candidates (YAu bcc(100) for H, ZnPt bct(101) for O, and CdAu bcc(101) for OH). First, for the case of YAu bcc(100) found for the adsorption of H, the sample of the dataset from which the counterfactual was generated was ScCu bcc(101) with an E_{ads} of −0.37 eV. The counterfactual that led to the discovery of the YAu surface involved a decrease in Ψ (from 29.94 to 18.98) and an increase in WEN (from 1.63 to 1.72). Lowering Ψ to reduce the value of the E_{ads} was already discussed in relation to Fig. 3c, but the impact of the WEN on the adsorption of H was still unknown. Taking a look at the scatter plot in Fig. 5a, it can be observed that, in regions close to Ψ equal to 30, large values of WEN lead to less positive SHAP values (*i.e.*, less positive E_{ads}), while the trend is the opposite when Ψ is close to 20 with values of WEN greater than 1.8 presenting the most positive values. Further below a value of Ψ of 20, any correlation between the value of WEN and the adsorption energy seems to disappear. This confirms why the model suggests that a reduction of Ψ together with a slight increase in the WEN leads to a more negative E_{ads}. In the case of

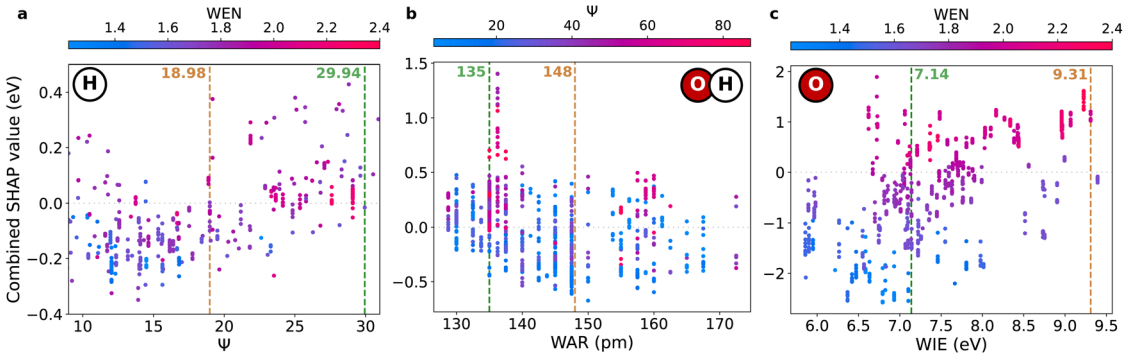

Fig. 5 Scatter plots of combined SHAP values. The data are plotted as a function of (a) Ψ values with points colored according to their WEN for H, (b) WAR with points colored according to their Ψ values for OH, and (c) WIE with points colored according to their WEN for O. All samples in the dataset were used to make these plots. For (a), only the range of Ψ values from 9 to 31 was considered for the plot for the sake of clarity. The combined SHAP values were calculated as the sum of the SHAP values for the two variables considered (*e.g.*, WIE and WEN for O in (c)).

CdAu bcc(101) found for the adsorption of OH, the original sample was ZnAu bcc(101) with an E_{ads} of 1.36 eV. The generated counterfactual from this sample was a decrease in Ψ (from 60.54 to 50.08) and an increment in the WAR (from 135 to 148 pm). Similar to the case of H, the importance of Ψ on the E_{ads} was already found by the SHAP analysis (Fig. 3), while the WAR was not among the most important features. Fig. 5b shows a scatter plot relating the WAR and Ψ with the impact on the predicted E_{ads} in terms of the SHAP values. From this plot, it is evident that for surfaces with WARs between 135 and 140 pm, adsorption sites with values of Ψ above 50 are quite more unfavorable than those with smaller values. Moreover, when moving from this region to values of the WAR from 140 up to 150 pm we observe, in general, lower SHAP values. These conclusions allow us to understand the result of the counterfactual. It found a less positive E_{ads} by reducing the value of Ψ and increasing the WAR of the surface. Finally, a more complicated counterfactual led to the discovery of ZnPt bct(101) for the adsorption of O. The original sample was DyPd bcc(101) with a very negative E_{ads} (−5.15 eV). The generated counterfactual modified four features: WEN (from 1.71 to 2.03), Ψ (from 85.87 to 72.68), WAR (from 157.5 to 135 pm), and WIE (from 7.14 to 9.31 eV). The effects of increasing WEN were already discussed in the context of the SHAP analysis (Fig. 3). Larger electronegativities lead to a less favorable adsorption for O. In the case of Ψ, although the feature is not as important as for the other adsorbates, if we focus in Fig. 3c on values of $(\prod_{i=1}^{N} S_i)^{\frac{2}{N}}$ around 120 (a site with a 1 : 1 proportion of the elements will have this value, since Dy has 12 and Pd has 10 outer electrons) we observe a strong negative correlation between Ψ values and the electronegativity of the site. Hence, smaller values of Ψ are also linked to less negative E_{ads}. The role of the WAR is similar to the one discussed for OH (the trends found in Fig. 5b are analogous to those for O in Fig. S5 in the SI). WAR values between 135 and 140 pm present very positive SHAP values for adsorption sites with values of Ψ higher than 60. The last feature, WIE, is the one for which we do not yet have any insight. In Fig. 5c, we show a scatter plot similar to previous ones where we show the SHAP values for the different combinations of WIE and WEN in the dataset. As expected, the surfaces with the largest WEN values show the most positive SHAP values regardless of their WIE. However, there is an evident increase in SHAP values for WIEs above 9 eV that correspond to less favorable E_{ads}. The analysis of these rather small changes enables an understanding of why a surface with such characteristics presents a considerable increase in E_{ads} compared to DyPd bcc(101).

During the process of retrieval of candidates (Fig. 1b), the features found by the generation of counterfactuals are modified in order to fit existing materials, surfaces, and adsorption sites. However, we can see from the right-hand side of Fig. 5 that they do not differ much from their counterfactual values. Indeed, except for the values of Ψ of YAu bcc(100) and CdAu bcc(101) that deviate 43% and 20%, respectively, all features change by less than 8% with respect to their counterfactuals. Moreover, with the insights obtained from analyzing the counterfactual explanations, it is easy to understand why the E_{ads} is as different as it is with respect to the E_{ads} predicted from the counterfactual features (and, hence, from the original sample). For YAu bcc(100), a value of Ψ of 10.83 is an indicator of a more negative E_{ads} than the original value of 29.94 of the longbridge-Cu site at ScCu bcc(101), but of a slightly less negative one than that of the counterfactual as can be observed in Fig. 5a when comparing the general behavior of points in the ranges of Ψ from 10 to 12 and from 12 to ∼18. Also, for Ψ values around 10, WENs from ∼1.8 (as the one of YAu) might lead to more positive E_{ads}. In the case of ZnPt bct(101), there are two changes that point towards a more negative E_{ads} compared to the one of the counterfactual. Both the WEN and the WIE decrease. However, these changes are rather small (0.1 for WEN and 0.13 eV for WIE) and, therefore, it has a very similar E_{ads} as that of the counterfactual. Finally, the value of Ψ of the longbridge-Au site at CdAu bcc(101) is almost the same as the original value of the longbridge-Au site at the ZnAu bcc(101) surface. Hence, most of the decrease of E_{ads} can be attributed solely to the large WAR of the surface (10 pm larger than in ZnAu). The less positive E_{ads} with respect to the counterfactual despite Ψ being larger (less favorable for adsorption; see Fig. 3a)

can be linked to a change in WIE from 9.31 eV (for both ZnAu and the counterfactual) to 9.11 eV for CdAu since the impact of WIE on the E_{ads} of OH shows similar trends to that on the E_{ads} of O (the plot for OH can be found in Fig. S6 in the SI for comparison).

3 Conclusions

We presented a novel strategy for the discovery and design of new materials based on XAI. Specifically, we used counterfactual explanations to get a set of features as close as possible to those of a known sample in the dataset but with a desired target value of the property of interest. Explainability is ensured by construction since every candidate found can be linked to the original sample from which the counterfactual was generated. As important steps before a discovery and design task, we highlighted the need for both a typical model selection process and validation through chemical and/or physical concepts.

On the example of heterogeneous catalysts for hydrogen production and energy generation, we trained ML models for the prediction of E_{ads} of H, O, and OH on binary intermetallic compounds (A_xB_y) and pure metals. ET models resulted in being the most accurate for the prediction of the target property for all the considered adsorbates, showing both a good global performance (in terms of MAE, RMSE, and R^2) and a low rate of outlier predictions. SHAP analysis of the ET models further confirmed their reliability by demonstrating that their predictions are obtained by leveraging chemical and physical concepts. Local environment features, such as the GCN, that are strongly related to the size of the adsorption site, or Ψ, which describes the chemical environment of the adsorption site through the electronic properties of the nearest neighbors, resulted in being the determinant for the prediction of E_{ads}.

The coupling of our ML models with DiCE for generating the counterfactuals within the XAI strategy offered more than 1000 candidates per adsorbate that were reduced to 8 after the candidates' retrieval and subsequent validation. Among these discovered candidates, 4 of them were confirmed through reference DFT calculations: YAu bcc(100), ZnIr hcp(10−11), ZnPt bct(101), and CdAu bcc(101). Then, explanations to gain insights into why the discovered surfaces are better than others were devised by comparing original samples, counterfactuals, and discovered candidates. Such explanations allowed us to unveil subtle relationships between the most relevant features (as found in SHAP analysis), other, in principle, less important features, and the E_{ads}. For instance, the WAR and the WIE were found to be determinants in tuning the E_{ads} to reach the target values.

The most intricate, oftentimes non-intuitive, correlations are hard to devise not only for human experts, but also with basic explanation tools (such as global feature analysis). However, as a well-trained ML model is capable of exploiting them effectively, all the insights and knowledge can be recovered if an appropriate XAI tool is employed within the ML pipeline. Since AI-assisted discovery of molecules and materials is meant to revolutionize several fields (*e.g.*, medicine, energy production, and industry, among others) by designing new potential compounds with specific properties, the implementation of XAI for these approaches is a logical next step in their development. Designing through AI should not be only about discovering new molecules and materials, but also about understanding why and how a formerly unknown compound is good, or even better, for certain application(s) than the currently used ones. These insights can implicitly lead to the creation of "recipes" of how to modify a molecule or a material in order to improve or deplete a target property. In the present study, we have worked towards that direction by developing an alternative to already available discovery and design strategies, such as high-throughput screening or generative models, but that for the first time incorporates explainability as its main driving mechanism. Given the demonstrated reach of our strategy, we expect that, in the near future, most of the ML endeavors in computational chemistry will incorporate XAI in one of its forms to achieve accurate and reliable ML models that can provide knowledge leading to new insights into chemistry and physics.

Author contributions

V. V.-G. conceived and designed the idea behind this work and performed the computational studies and the algorithmic development. Both authors were involved in the analysis, interpretation, discussion of the results, and the writing of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Python routines to perform each of the steps of the workflow shown in Fig. 1 are publicly available *via* GitHub at **https://github.com/vvassilevg/HighHydrogenML-XAI**.

Most data used to make the figures are available in the supplementary material as independent files. Those figures for which the data are not directly available can be easily created from the publicly available Python routines.

Although the datasets used to train the ML models in this work were obtained from ref. 49, we provide, in the supplementary material, all the datasets since, for this work, we split the original O–OH dataset into two independent datasets and changed slightly the format of the files for fitting our Python routines.

All models used for performing the XAI discovery and design task (ET models trained on all samples in the dataset) are available in the supplementary material. All other models discussed in this work can be easily obtained with the publicly available Python routines since all training, validation, and test splits are provided in the supplementary material.

A file in xlsx format is provided in the supplementary material containing the information of original samples, counterfactuals, and validated candidates.

Supplementary information (SI) is available. See DOI: **https://doi.org/10.1039/d5sc06442b**.

Acknowledgements

VV-G acknowledges support through the project HighHydrogenML (GA number 101105610) funded by the Horizon Europe program of the European Union. This investigation was also partly supported by the project "High-throughput strategies for the discovery of new catalysts for the hydrogen economy through elastic strain engineering" (CATbyESE), funded in the call for Oriented Projects for the Ecological and Digital Transition, Spanish Ministry of Science and Innovation (TED2021-129497B-I00), and by the project "Digital strategies for autonomous discovery of materials for engineering applications" (DIGIMATER-CM, reference TEC-2024/TEC-102), funded in the call of "Programas de Actividades de I + D" of the Comunidad de Madrid. Computational resources and technical assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa) are gratefully acknowledged.

Notes and references

1 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Muller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.
2 G. B. Goh, N. O. Hodas and A. Vishnu, *J. Comput. Chem.*, 2017, **38**, 1291–1307.
3 R. Ramakrishnan and O. A. von Lilienfeld, *Rev. Comput. Chem.*, 2017, **30**, 225–256.
4 N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. S. Smith, B. Nebgen, R. Messerly, Y. W. Li, A. I. Boldyrev, K. Barros, *et al.*, *Nat. Rev. Chem.*, 2022, **6**, 653–672.
5 A. Sabagh Moeini, F. Shariatmadar Tehrani and A. Naeimi-Sadigh, *Sci. Rep.*, 2024, **14**, 26736.
6 T. Wang, K. Zhang, J. Thé and H. Yu, *Comput. Mater. Sci.*, 2022, **201**, 110899.
7 A. K. Gupta and K. Raghavachari, *J. Chem. Theory Comput.*, 2022, **18**, 2132–2143.
8 S. Zhang, M. Chigaev, O. Isayev, R. A. Messerly and N. Lubbers, *J. Chem. Inf. Model.*, 2025, **65**, 4367–4380.
9 L. Cao, *Trends Chem.*, 2022, **4**, 347–360.
10 G. Yuan, M. Wu and L. Ruiz Pestana, *J. Phys. Chem. C*, 2023, **127**, 15809–15818.
11 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schutt, A. Tkatchenko and K.-R. Muller, *Chem. Rev.*, 2021, **121**, 10142–10186.
12 D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdau, D. J. Cole, *et al.*, *J. Am. Chem. Soc.*, 2025, **147**, 17598–17611.
13 A. Kabylda, J. T. Frank, S. S. Dou, A. Khabibrakhmanov, L. M. Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller and A. Tkatchenko, *ChemRxiv*, 2025, preprint, DOI: **10.1021/jacs.5c09558**.
14 B. Ryu, L. Wang, H. Pu, M. K. Chan and J. Chen, *Chem. Soc. Rev.*, 2022, **51**, 1899–1925.
15 Z. Rao, P.-Y. Tung, R. Xie, Y. Wei, H. Zhang, A. Ferrari, T. Klaver, F. Körmann, P. T. Sukumar, A. Kwiatkowski da Silva, *et al.*, *Science*, 2022, **378**, 78–85.
16 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
17 X. Yang, Y. Wang, R. Byrne, G. Schneider and S. Yang, *Chem. Rev.*, 2019, **119**, 10520–10594.
18 M. Staszak, K. Staszak, K. Wieszczycka, A. Bajek, K. Roszkowski and B. Tylkowski, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1568.
19 E. Ogoshi, M. Popolin-Neto, C. M. Acosta, G. M. Nascimento, J. N. Rodrigues, O. N. Oliveira Jr, F. V. Paulovich and G. M. Dalpian, *Discover Mater.*, 2024, **4**, 6.
20 B. Kang, C. Seok and J. Lee, *J. Chem. Inf. Model.*, 2020, **60**, 5984–5994.
21 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, *Chem*, 2020, **6**, 3100–3117.
22 M. Ancona, E. Ceolini, C. Öztireli and M. Gross, *arXiv*, 2017, preprint, arXiv:1171.06104, DOI: **10.48550/arXiv.1171.06104**.
23 F. Oviedo, J. L. Ferres, T. Buonassisi and K. T. Butler, *Acc. Mater. Res.*, 2022, **3**, 597–607.
24 X. Wang, A. S. Barnard and S. Li, *Intell. Comput.*, 2025, **4**, 0126.
25 M. Gallegos, V. Vassilev-Galindo, I. Poltavsky, Á. Martín Pendás and A. Tkatchenko, *Nat. Commun.*, 2024, **15**, 4345.
26 X. Chen, D. Yu, L. Zhao and F. Liu, *Digital Discovery*, 2025, **4**, 2062–2074.
27 J. Jiménez-Luna, M. Skalic, N. Weskamp and G. Schneider, *J. Chem. Inf. Model.*, 2021, **61**, 1083–1094.
28 K. Chen, G. Chen, J. Li, Y. Huang, E. Wang, T. Hou and P.-A. Heng, *J. Cheminf.*, 2023, **15**, 43.
29 K. Morita, D. W. Davies, K. T. Butler and A. Walsh, *J. Chem. Phys.*, 2020, **153**, 024503.
30 V. V. Korolev, A. Mitrofanov, E. I. Marchenko, N. N. Eremin, V. Tkachenko and S. N. Kalmykov, *Chem. Mater.*, 2020, **32**, 7822–7831.
31 S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson and C. Shah, *ACM Comput. Surv.*, 2024, **56**, 1–42.
32 G. P. Wellawatte, A. Seshadri and A. D. White, *Chem. Sci.*, 2022, **13**, 3697–3705.
33 A. Lamens and J. Bajorath, *RSC Med. Chem.*, 2024, **15**, 1547–1555.
34 G. P. Wellawatte, H. A. Gandhi, A. Seshadri and A. D. White, *J. Chem. Theory Comput.*, 2023, **19**, 2149–2160.
35 D. Numeroso and D. Bacciu, *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.
36 A. Lamens and J. Bajorath, *ChemMedChem*, 2024, **19**, e202300586.
37 D. W. Robbins and J. F. Hartwig, *Science*, 2011, **333**, 1423–1427.
38 O. Schilter, A. Vaucher, P. Schwaller and T. Laino, *Digital Discovery*, 2023, **2**, 728–735.
39 M. Popova, O. Isayev and A. Tropsha, *Sci. Adv.*, 2018, **4**, eaap7885.
40 A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, *Mol. Pharm.*, 2017, **14**, 3098–3104.
41 F. Dinic, I. Neporozhnii and O. Voznyy, *Comput. Mater. Sci.*, 2024, **231**, 112580.
42 P. Sabatier, *La catalyse en chimie organique*, ed. C. Béranger, 1920, vol. 3.

43 V. Viswanathan, H. A. Hansen, J. Rossmeisl and J. K. Nørskov, *ACS Catal.*, 2012, **2**, 1654–1660.
44 J. Xie, H. Qu, J. Xin, X. Zhang, G. Cui, X. Zhang, J. Bao, B. Tang and Y. Xie, *Nano Res.*, 2017, **10**, 1178–1188.
45 M. K. Sabbe, M.-F. Reyniers and K. Reuter, *Catal. Sci. Technol.*, 2012, **2**, 2010–2024.
46 A. Kiejna, G. Kresse, J. Rogal, A. De Sarkar, K. Reuter and M. Scheffler, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 035404.
47 V. Butera, *Phys. Chem. Chem. Phys.*, 2024, **26**, 7950–7970.
48 G. Kresse and D. Joubert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 1758.
49 C. Martínez-Alonso, V. Vassilev-Galindo, B. M. Comer, F. Abild-Pedersen, K. T. Winther and J. LLorca, *Catal. Sci. Technol.*, 2024, **14**, 3784–3799.
50 F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet and D. Loffreda, *Angew. Chem., Int. Ed.*, 2014, **53**, 8316–8319.
51 W. Gao, Y. Chen, B. Li, S.-P. Liu, X. Liu and Q. Jiang, *Nat. Commun.*, 2020, **11**, 1196.
52 S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1339–1345.
53 X. Ma, Z. Li, L. E. Achenie and H. Xin, *J. Phys. Chem. Lett.*, 2015, **6**, 3528–3533.
54 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.
55 M. Evans and M. Polanyi, *Trans. Faraday Soc.*, 1936, **32**, 1333–1360.
56 L. S. Shapley, *Proc. Natl. Acad. Sci. U. S. A.*, 1953, **39**, 1095–1100.
57 S. M. Lundberg and S.-I. Lee, *Adv. Neural Inf. Process Syst. 30*, Curran Associates, Inc., 2017, pp. 4765–4774.
58 S. Wachter, B. Mittelstadt and C. Russell, *Harv. J. Law Technol.*, 2017, **31**, 841.
59 R. K. Mothilal, A. Sharma and C. Tan, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.
60 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, *APL Mater.*, 2013, **1**, 011002.
61 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.
62 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
63 T. Chen and C. Guestrin, *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
64 F. Chollet, *et al.*, *Keras*, 2015, **https://keras.io**.
65 T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, *KerasTuner*, 2019, **https://github.com/keras-team/keras-tuner**.
66 N. İnoğlu and J. R. Kitchin, *Mol. Simul.*, 2010, **36**, 633–638.
67 B. Hammer and J. K. Nørskov, *Adv. Catal.*, 2000, **45**, 71–129.