Reaction Chemistry & Engineering

ROYAL SOCIETY OF CHEMISTRY

PAPER

Published on 16 July 2026

Cite this: DOI: 10.1039/d6re00180g

Bridging structure and activity in nanocatalysts *via* machine learning and global structure representations

Sofia Zinzani, [a] Francesca Baletto [a] and Kevin Rossi *[bc]

Received 16th May 2026,
Accepted 14th July 2026

DOI: 10.1039/d6re00180g

rsc.li/reaction-engineering

Establishing a mapping between nanocatalysts structure and their catalytic properties is essential for efficient design. To this end, we develop a general machine learning framework, validated on a representative and challenging application: predicting the mass activity of Pt nanoparticles for the electrochemical oxygen reduction reaction, estimated *via* a microkinetic model. Accurate models are obtained when leveraging either a nanocatalyst's structure representation accessible at the computational level, namely the surface site generalized coordination number distributions, or one accessible experimentally, namely the nanoparticle's pair distance distribution function. The same representations also enable to predict with a good accuracy the stability of the same nanoparticles. Building on this result, we demonstrate that our machine learning approach, in tandem with multi-objective Bayesian optimization, efficiently identifies with a large pool, candidate nanocatalysts that display an optimal trade-off between energetic stability and activity. These findings provide a robust blueprint for accelerated theoretical and experimental identification of active nanocatalysts.

1 Introduction

The development of next-generation heterogeneous catalysts with enhanced activity, selectivity, and stability underpins the large-scale deployment of clean energy solutions, the production of value-added chemicals, and the mitigation of carbon emissions.[1] Indeed global efforts towards chemical processes which are sustainable, cost-competitive, and display less at risk dependencies require transformative advances in the design of heterogeneous catalysts, which currently appear in at least one step of 80% industrial processes.[2] In this context, nanocatalysts provide a versatile platform for tuning catalytic performance through controlled manipulation of their size, shape, chemical composition and ordering.[3,4] Such flexibility allows researchers to optimize surface electronic structures, surface sites adsorption properties, and active site density, to boost reactivity and selectivity in a wide range of applications.

Machine learning (ML) models can learn intricate structure–reactivity correlations from simulations or experiments, and efficiently provide surrogate performance predictions for new structures or compositions.[5,6] Indeed ML applications have obtained remarkable successes in predicting intermediate-scale properties, including adsorption energies and reaction barriers, which are crucial descriptors in catalytic processes.[7–9] A gap still exists in reliably establishing a direct map between morphological features of catalysts and macroscopic performance metrics – such as the overall catalytic activity.[10]

In this work, we demonstrate a general ML approach capable to address this shortcoming. As an illustrative case study, we focus on learning structure–property relationship in platinum (Pt) nanoparticles (NPs) for electrochemical oxygen reduction reaction (eORR) – a key process in green energy technologies such as fuel cells and electrolyzers.[11] Given the high cost and limited availability of Pt, it is indeed desirable to identify nanocatalyst geometries that maximize catalytic performance per unit mass. The ORR activity of Pt is known to depend on surface structure, with high-index facets often exhibiting enhanced catalytic performance compared with low-index planes because of their higher densities of steps, edges, and kink sites.[12] At the same time, optimizing catalytic activity per Pt mass requires minimizing particle size while preserving sufficiently developed facets. In this regard, past experimental and theoretical studies identified an optimal Pt nanoparticle size sin the range of approximately 2.5–3.5 nm (ref. 13–18) and limited gains in per-site activity above ~5 nm. Reported mass activity reach as high as to times reference Tanaka commercial Pt/C catalyst, corresponding to a value just below 1 A mg^{-1} at applied potential close to 0.9 V, for low concentrations of weakly adsorbing electrolytes (*e.g.*, 0.1 M $HClO_4$).

We validate the approach by considering a descriptor of the NP surface that is accessible at the computational level and known for its capability to predict catalytic activity in Pt NPs for

[a] *Physics Department, Università degli Studi di Milano, Milano, Italy*
[b] *Materials Science and Engineering Department, TU Delft, Delft, Netherlands*
[c] *Climate Safety and Security Centre, TU Delft, The Hague, Netherlands.*
E-mail: k.r.rossi@tudelft.nl

eORR: the GCN distribution of atop sites.[19,20] We, then, illustrate the method generality and effectiveness considering an experimentally accessible representation of the nanosystem, namely its pair distance distribution function (PDDF).[21,22]

Finally, we consider not only structure–activity relationship, but also (simplified) structure–stability metrics, and show how a multi-objective Bayesian optimization[23] approach enables to screen a large pool of candidate structures to efficiently identify the ones presenting best trade-offs between activity and stability.

While the present framework relies on simplified approaches to estimate ground truth catalyst activity and stability, our results provide evidence that compact structural representations encode key information governing catalyst performance. Further, the generality of the approach suggests a broad applicability across diverse materials and reactions, towards accelerated nanocatalyst discovery.

2 Methods

2.1 Structure generation

In line with size-dependent trends reported in the literature, we consider a dataset comprising 52 643 unique nanoparticle structures, with sizes ranging from 264 to 2830 atoms, corresponding approximately to diameters between 2 and 5 nm.

To generate the dataset, we collected configurations sampled along coalescence trajectories,[24,25] thereby capturing a broad variety of non-ideal and structurally diverse geometries beyond highly symmetric model nanoparticles. All sampled structures were then treated as monometallic Pt NPs. To increase both the structural and size diversity of the dataset, we additionally applied a systematic surface-modification procedure in which a random number of surface atoms was removed from each nanoparticle. Following these modifications, all structures were relaxed using interatomic interactions described within the second-moment tight-binding approximation (see SI section 1 for additional detail). The relaxation step ensures that the final dataset contains (meta)stable structure, weakly correlated with the original configurations sampled during the coalescence trajectories.

For each relaxed structure, we record ΔE, *i.e.* the energy in excess with respect to having N atoms in the bulk, weighted by the number of atoms in the surface:[26,27]

$$\Delta E = \frac{N E_{\mathrm{coh}} - E_{\mathrm{pot}}}{N^{2/3}} \tag{1}$$

where N is the number of atoms in the cluster and E_{coh} is the cohesive energy. A $\Delta E = 0$ corresponds to a configuration as stable as in the bulk. The more positive the ΔE, the more it is likely that the structure under consideration may undergo phase change into a more stable configuration. We emphasize that ΔE should be interpreted as a descriptor of the intrinsic thermodynamic stability of the nanoparticle structure. While it provides information on the energetic cost associated with finite-size and surface effects, it does not explicitly account for degradation mechanisms occurring under reaction conditions, such as surface poisoning, oxidation, dissolution, reconstruction, or sintering. Consequently, ΔE is used here as a computationally tractable proxy for stability rather than as a quantitative predictor of catalyst durability under operando electrochemical conditions.

2.2 Mass activity estimate

To estimate nanoparticles mass activity for eORR at an applied voltage of 0.9 V ($\mathrm{MA}_{@0.9\mathrm{V}}$), we adopt a microkinetic model we previously developed and validated against experimental data.[20] In the latter, surface site contributions to the overall nanoparticle activity for eORR are estimated as a function of their atop generalized coordination number (GCN)[28,29] for a site j, defined as:

$$\mathrm{GCN}_j = \frac{\sum_{i \in NN_j} \mathrm{CN}_i}{\mathrm{CN}_{\max}} \tag{2}$$

where $\mathrm{CN}_{\max}$ is equal to the bulk coordination number ($\mathrm{CN}_{\max} = 12$ for fcc metals), and CN_i labels the coordination numbers of the atom i, which is a nearest neighbour of the atom j. The coordination number (CN), in turn, reads:

$$\mathrm{CN}_{\mathrm{tot}} = \sum_{\substack{i,j \\ i \neq j}} f(r_{ij}), \tag{3}$$

$$f(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} \leq d_0, \\ 0 & \text{if } r_{ij} > d_0, \end{cases} \tag{4}$$

The cut-off distance d_0 is defined as the first minimum of the pair distance distribution function (PDDF). For simplicity, we opt to use a fixed cut-off value equal to 0.85× the lattice parameter. We compute both the coordination number and the atop GCNs using an in-house open-source python code.[30]

Next, we label the collection of α GCNs values across the non-equivalent surface sites in the nanoparticle as $\mathscr{T}_{\mathrm{isomer}}$. We then estimate the total current density for eORR, under an applied electrochemical potential U, at a temperature T, as:

$$j_{\mathrm{isomer}}(U,T) = \sum_{\alpha \in \mathscr{T}_{\mathrm{isomer}}} \frac{\Omega(\alpha)}{N_{\mathrm{site}}} C\alpha \exp(-\beta \Delta G(\alpha, U)) \tag{5}$$

where, $\beta = 1/k_{\mathrm{B}}T$ is the Maxwell–Boltzmann factor, and $\Omega(\alpha)$ denotes the occurrence of adsorption sites with GCN α, and N_{site} is the total number of available sites at the nanoparticle surface. We consider $T = 300$ K and $U = U^* = 0.9$ *V*. The pre-factor $C = 12$ was calibrated to reproduce known current densities for low-Miller-index surfaces of reference Pt catalysing electrochemical oxygen reduction reaction in weakly adsorbed electrolytes. The free-reaction energy of the reaction step $\Delta G(\alpha, U)$ depends on both the local geometry.

To describe the variation $\Delta G(\alpha, U)$ for OH* adsorption in terms of adsorption site coordination (under the assumption that this is the rate limiting step in the reaction), we employ a volcano-type relationship established in the literature:

$$\Delta G_{OH}(\alpha) = \begin{cases} -0.724 + 0.192\alpha & \text{for } \alpha < 8.3 \\ -2.345 - 0.178\alpha & \text{for } \alpha > 8.3 \end{cases} \quad (6)$$

The latter was validated against density functional theory data, and resulting trends were further validated against experiments.[31] For reference, in SI – Fig. S1 we report the volcano plot and correspond site specific currents.

Following the computational hydrogen electrode model approach, the electrochemical potential is introduced as an *a posteriori* correction to the free energy estimate.

The specific activity of the isomer, j^*_{isomer} (in mA cm^{-2}), is computed by evaluating eqn (5) at a given operating bias U^*. The corresponding mass activity is defined as:

$$MA^*_{isomer} = \frac{j^*_{isomer} \cdot A_{isomer}}{M_{isomer}} \quad (7)$$

Here, M_{isomer} is the total mass of the nanoparticle, equal to the number of Pt atoms times the atomic mass (195 a.u.), while A_{isomer} is the effective surface area, approximated by:

$$A_{isomer} = \sum_{i=1}^{N} 4\pi r_{at}^2 \left(1 - \frac{GCN(i)}{12}\right) \quad (8)$$

where the sum is over all the atoms i in the nanoparticle with generalized coordination lesser than 11.

2.3 Structural representations and featurization

The choice of structural representation plays a crucial role in machine learning applications, as it determines how the relevant physicochemical information is encoded and ultimately accessible to the predictive model.

Here, we employ two complementary approaches to represent nanoparticle structures: (i) the Generalized Coordination Number (GCN) distributions, and (ii) the Pair Distance Distribution Function (PDDF). For the GCN representation, the feature vector is constructed from the distribution of surface-site GCN values, where each feature corresponds to the population associated with a given GCN interval. For the PDDF representation, the feature vector is instead obtained from the histogram of interatomic distances, with each feature representing the occurrence frequency of atom pairs within a specific distance bin. The workflow adopted to evaluate these representations is illustrated in Fig. 1 and 2, respectively.

2.4 Gaussian process regression

The Gaussian process regression (GPR) used for learning structure–property and structure–stability relationships employs a composite kernel consisting of a constant kernel $C(c)$ multiplied by a Matern kernel, with an additional white noise component:

$$k(\mathbf{x}_i, \mathbf{x}_j) = c \cdot k_{Matern}(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2 \delta_{ij} \quad (9)$$

where the Matern kernel with smoothness parameter $\nu = 1.5$ is defined as

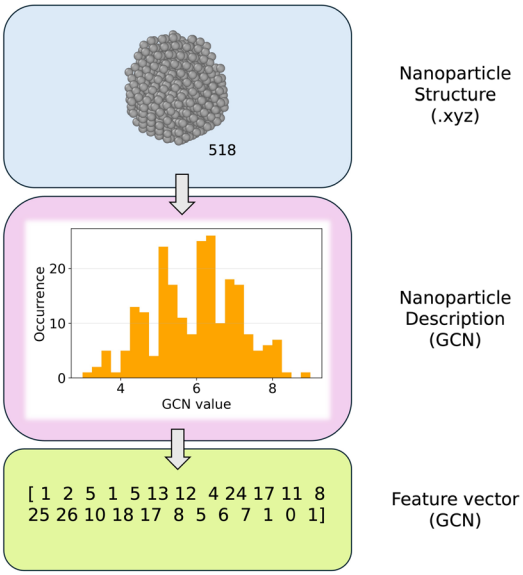

Fig. 1 Top) a snapshot of one Pt nanocatalyst of 518 atoms. Mid) its GCN distribution. Bottom) the associated feature vector used as input in the ML workflow.

$$k_{Matern}(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{\sqrt{3}||\mathbf{x}_i - \mathbf{x}_j||}{\ell}\right) \exp\left(-\frac{\sqrt{3}||\mathbf{x}_i - \mathbf{x}_j||}{\ell}\right) \quad (10)$$

where c is the constant scaling factor initialized at 1.0, ℓ is the length-scale parameter initialized at 1.0, ν = 1.5 controls the smoothness of the process, and σ_n^2 represents the white noise variance initialized at 10^{-3}.

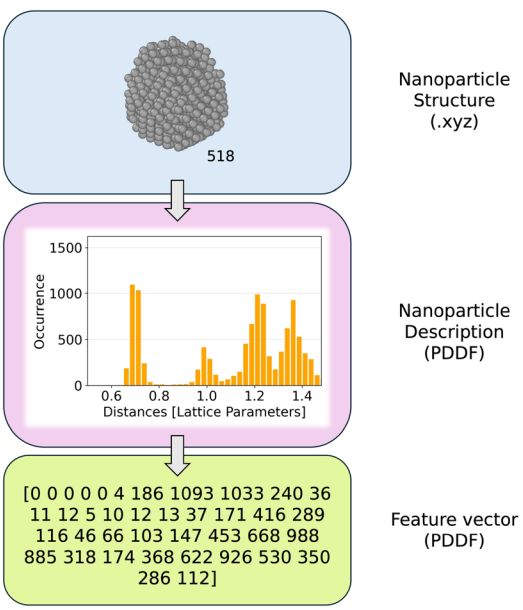

Fig. 2 Top) a snapshot of one Pt nanocatalyst of 518 atoms. Mid) its PDDF distribution. Bottom) the associated feature vector used as input in the ML workflow.

The GPR predictions are given by:

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^T(\mathbf{K} + \alpha\mathbf{I})^{-1}\mathbf{y}$$
$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \alpha\mathbf{I})^{-1}\mathbf{k}_* \quad (11)$$

where $\mathbf{K}$ is the kernel matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_*$ is the vector of kernel evaluations between the test point and training points, $\alpha = 10^{-6}$ is the noise level for numerical stability, $\mathbf{I}$ is the identity matrix and $\mathbf{x}_i$ denotes the feature vector corresponding to the i-th sample (*i.e.*, the GCN or PDDF distributions as discussed in section 2.3 and illustrated in Fig. 1 and 2).

GPR fit and prediction are performed using sklearn[32] library.

2.5 Multi-objective Bayesian optimization

Our multi-objective Bayesian optimization (MOBO)[23] framework utilizes the predicted activity from Gaussian process regression and its uncertainty to guide candidate selection *via* Pareto front upper confidence bound[33] acquisition function, to balance exploration and exploitation.

We fix a initial training set size to 2048 points. A Gaussian process model is trained on these data and used to predict both the $MA_{@0.9V}$ and Δ (where $\Delta = -\Delta E$) and their associated uncertainties, over the remaining candidate pool. The Pareto front of $MA_{@0.9V}$ *vs.* Δ optima is then computed in the acquisition space defined by the upper confidence bounds. The subset of non-dominated candidates is selected. The selected points are then evaluated, added to the training set, and used to update the model iteratively, progressively refining the model and improving the identification of optimal regions in the multi-objective landscape. If the resulting Pareto set exceeds a predefined batch size (train_pool_number = 100), candidates are further ranked according to their acquisition values and the most informative samples are retained. For the multi-objective Bayesian optimization (MOBO) framework, a simpler kernel configuration was adopted, as it proved sufficient for the optimization task. In particular, an RBF kernel with Automatic Relevance Determination (ARD) was employed:

$$k_{\mathrm{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sum_{d=1}^{D}\frac{(x_{i,d} - x_{j,d})^2}{2\ell_d{}^2}\right) \quad (12)$$

where ℓ_d denotes the individual length-scale associated with the d-th input dimension through ARD. The multi-task Gaussian process model was implemented using a rank-1 task covariance structure with num_tasks equal to output tasks. MOBO runs are performed using GPY[34] library.

3 Results

3.1 Predicting structure–activity and structure–stability from local surface structural descriptors

As an initial proof of concept, we discuss the possibility of machine learning Pt nanoparticles MA for eORR, when adopting the GCN as the nanoparticle structure descriptor.

The initial database of 52 318 structure results in diverse GCN distributions (binned with a resolution 0.25) as illustrated by the histograms in Fig. 3. Because of our binning resolution, GCN signature at $6.5 < \mathrm{GCN} \leq 6.75$ and $7.25 < \mathrm{GCN} \leq 7.5$ correspond to the presence of (100) and (111) facets, respectively. Atoms with $\mathrm{GCN} > 8$ instead correspond to convex-sites, which are highly active. Previously, sites with GCN 8.3 were demonstrated to correspond to the most active Pt sites for eORR.[31]

The distribution of values of $MA_{@0.9V}$ for our dataset is reported in SI – Fig. S2. The mean mass activity in the dataset corresponds to 808.40 mA mg^{-1} with a standard deviation of 403.82 mA mg^{-1}, in line with values observed in comparable experiments.[35] Overall, the mass activities span between 0–3 A mg^{-1}, the latter value being larger than what commonly reported for Pt nanoparticles, but also lesser than what reported, *e.g.*, for tailored nanostructured Pt nanowires[36] at comparable conditions.

To analyse the relationship between model accuracy and training points samples, we evaluate learning curves, which we report in Fig. 3, top panel. The learning curve is estimated by recording the mean accuracy of six models trained with 2^X, $X = 3.13$ datapoints randomly extracted from the whole dataset. A significant improvement in the model accuracy is witnessed when the model is trained with at least 128 data points ($R^2 = 0.865 \pm 0.003$). The model accuracy saturates to a near-perfect accuracy, albeit with a slower pace, when accounting for additional training samples. For the case of 8192 data points, we observe a correlation coefficient of $R^2 = 0.9621 \pm 0.0004$ and a mean absolute error (MAE) = 54.7 ± 0.2 mA mg^{-1} between predicted and reference validation mass activity values (Fig. 3 middle panel, where training points are reported in red and validation point in blue).

Such a low MAE (which is one order of magnitude less than the typical mass activities considered in this work, and observed experimentally) is achievable and expected since the ground truth data are obtained *via* a microkinetic model that utilizes the catalyst surface sites GCN to estimate $MA_{@0.9V}$. Accordingly, the predictive accuracy of the model depends on the resolution used to discretize the GCN distribution (SI – Fig. S3). A finer binning retains more detailed structural information and generally improves prediction accuracy, whereas coarser binning reduces the amount of encoded structural detail and can lead to lower accuracy.

Illustration of the error distribution, 3 lower panel, shows that errors generally distributed around zero, together with slight model $MA_{@0.9V}$ underestimation for high $MA_{@0.9V}$ ground truths. Errors up to 10 times larger than the MAE (= 54.7 ± 0.2 mA mg^{-1}) are observed, yet, these are also accompanied by large uncertainties, demonstrating that the model is capable to providing reliable estimates of the domains within it may be trusted.

When considering structure–stability relationships, (Fig. 4) we observe similar learning trends. The model accuracy in predicting the NPs excess energies improves significantly when doubling the number of training datapoints, until 128

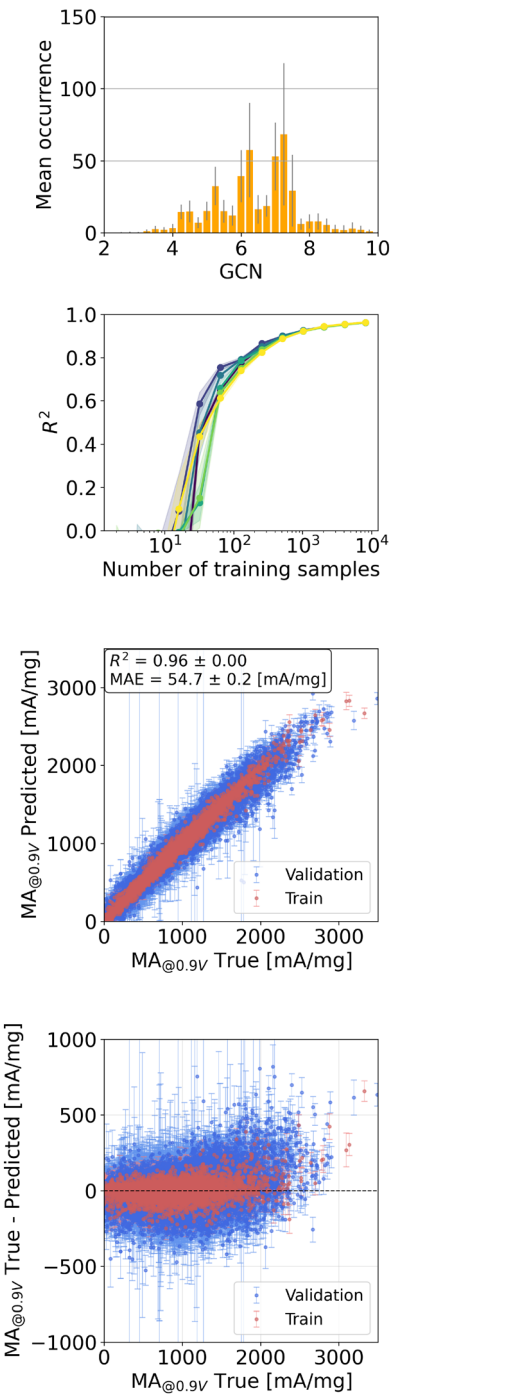

Fig. 3 $MA_{@0.9V}$ prediction from GCN – top) mean GCN distribution. To estimate this quantity we average the GCN distribution for each isomer in the dataset. The GCN distribution standard deviation, when considering each isomer in the dataset, is represented by the errorbars. GCN distribution bins have a resolution of 0.25. Middle-upper) GCN learning curve observed when training a machine learning model to predict Pt-NP activity for electrochemical oxygen reduction reaction from NP generalized coordination number distributions. Middle-lower) Corresponding parity plot between predicted and true activities for machine learning model trained with 8192 datapoints selected randomly. Bottom) Corresponding error distribution of the difference true – predicted activities with respect to the true values.

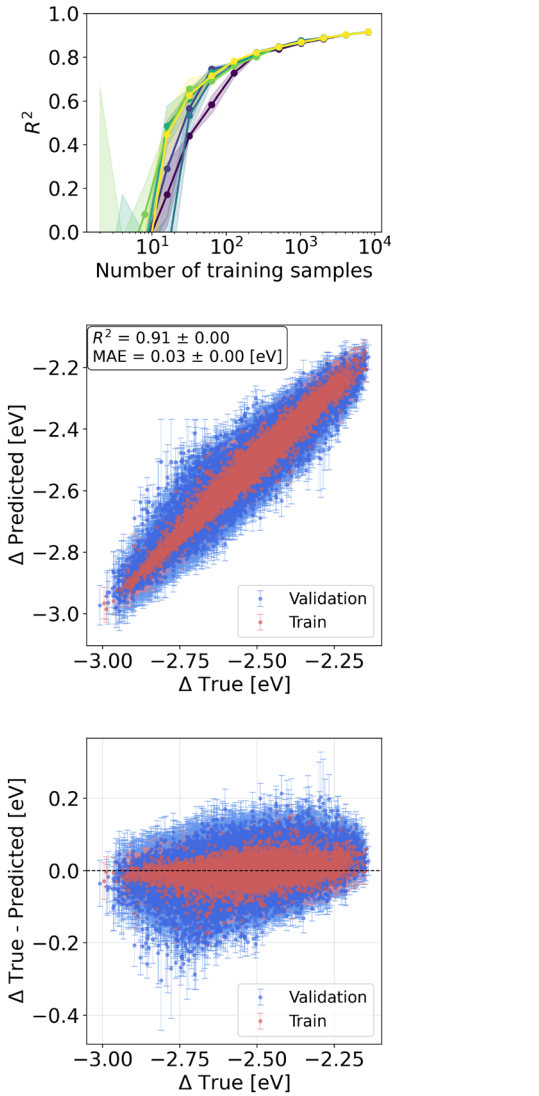

Fig. 4 Δ prediction from GCN – top) GCN learning curve observed when training a machine learning model to predict Pt nanoparticle stability for from the nanoparticle generalized coordination number distributions. Middle) Corresponding parity plot between predicted and true activities for machine learning model trained with 8192 datapoints selected *via* an active learning protocol. Bottom) Error distribution of the difference between true and predicted activities with respect to the true values.

configurations are accounted for. Diminishing returns in model accuracy are then observed (Fig. 4 top panel). For 8912 training datapoints (Fig. 4 – lower panel) we find a correlation coefficient of R^2 = 0.9149 ± 0.0001 and a mean absolute error close to 0.03 eV between predicted and reference excess energy values. The error distribution of the difference between true and predicted stability, is reported in the bottom panels of Fig. 4 and, similar to the case of $MA_{@0.9V}$ predictions, displays a fairly homogeneous arrangement.

Interestingly, despite the excess energy is a global property of the nanoparticle, we find that a descriptor based solely on the distribution of surface-site GCNs is sufficient to predict

nanoparticle stability with remarkably good accuracy. This result confirms that structural information governing stability may be effectively encoded within the surface atomic arrangement alone. Indeed, the surface atomic distribution is intrinsically correlated with the nanoparticle volume and naturally captures variations in surface energy.

3.2 Predicting structure–activity and structure–stability from a global structural descriptor

Building upon the promising results achieved using the GCN distribution as a global structural representation, we investigate an alternative one: the nanoparticle pair distance distribution function (PDDF).

The motivation for considering this representation lies in its proven connection to experimental observables. PDDFs can be obtained from total scattering measurements using X-rays, neutrons, or electrons.[37] In a previous work, we demonstrated how PDDFs track melting signatures and surface-bulk ordering in metallic nanoparticles.[22] More recently, the use of nanoparticle PDDFs to map structural heterogeneity in solid nanoparticles has been demonstrated too.[38] Further, experimentally, PDDF analyses of Pt-based nanostructures revealed that microstrain correlates with enhanced eORR kinetics.[39,40] Last but not least, approaches to reconstruct coordination number distributions from XAS spectra have been suggested.[21]

Our findings, summarized in Fig. 5, provide fair evidence supporting the hypothesis that machine learning models can effectively utilize the PDDF as a representation to predict Pt nanocatalysts' mass activity for eORR. We focus on interatomic pair distances up to a cut-off distance of 1.48 the bulk lattice parameter, *i.e.* accounting for variations in the first four neighbours peaks. In a crystalline FCC bulk, first and second nearest-neighbors, at $\sqrt{2}/2$ and 1 lattice distances, respectively, encode information about atoms first and second coordination shells. The four nearest neighbor distances are encoded when consider a cut-off distance of 1.48 the bulk lattice parameter. Deviation from ideal bulk lattice value in these distances, and beyond, encode bond distortions due to specific geometrical feature of the NP *e.g.*, grain boundaries, strain, and the presence of low coordination atoms or adatoms islands.

The learning curve depicted in Fig. 5 top panel demonstrates that a sizable but not intractable training set, comprising a few thousand samples, enables an ML model with a fair accuracy. A R^2 over 0.5 is indeed achieved with more than 1024 training samples. Notably, we do not observe a saturation in the model accuracy when increasing the number of training samples, within the training set sizes here considered, up to 8196 training samples. For the latter training size we observe a R^2 of 0.62 and a mean absolute error of 177.6 mA mg^{-1}, as shown in Fig. 5 middle panel. This value is still sizably smaller than mean mass activities observed for structures considered in this work, and more broadly in the literature.

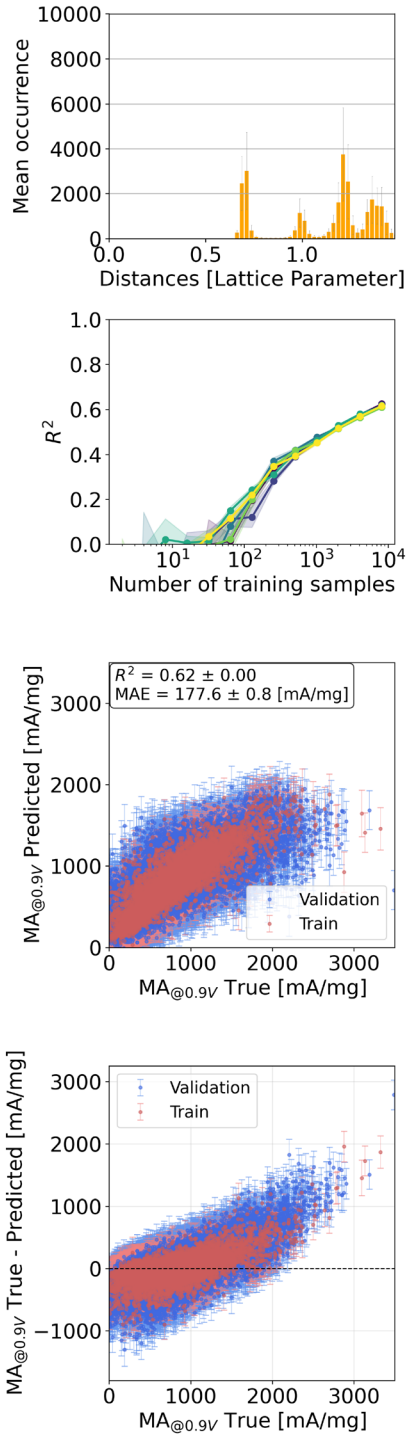

Fig. 5 $MA_{@0.9V}$ prediction from PDDF – top) mean PDDF values distribution over the full dataset of structures with its uncertainty represented by the errorbars. PDDF bins have a resolution of 0.01 lattice parameter. Middle-upper) PDDF learning curve observed when training a machine learning model to predict Pt nanoparticles activity for eORR from its PDDF. Middle-lower) Corresponding parity plot between predicted and true activities for machine learning model trained with 4020 datapoints selected *via* an active learning protocol. Bottom) Corresponding error distribution of the difference true – predicted activities with respect to the true values.

When probing the model error distribution, shown in the bottom panel of Fig. 5, we observe that the model understimates the activity of most active nanoparticles and overestimates the one of least active ones. While this behaviour is not desirable, we will discuss in the next section how such a shortcoming does not hinder model applications in the search for stable and active catalyst. Reported trends hold for different choices of pair distance distribution function binning and maximum distances considered (SI – Fig. S4 and S5).

When considering excess energy prediction, the PDDF emerges as an informative representation. Fig. 6 top panel show that model accuracy sharply increases (R2 *sim* 0.8) when accounting for at least 64 training points. Lower panel illustrate that. 8192 training points inform a model with a R2 of 0.98 and a MAE close to 0.02 eV. The accuracy of a PDDF based model can be rationalized in terms of two-body interactions providing leading contributions to the overall stability of metallic NPs, in agreement with results previously reported in the literature.[41]

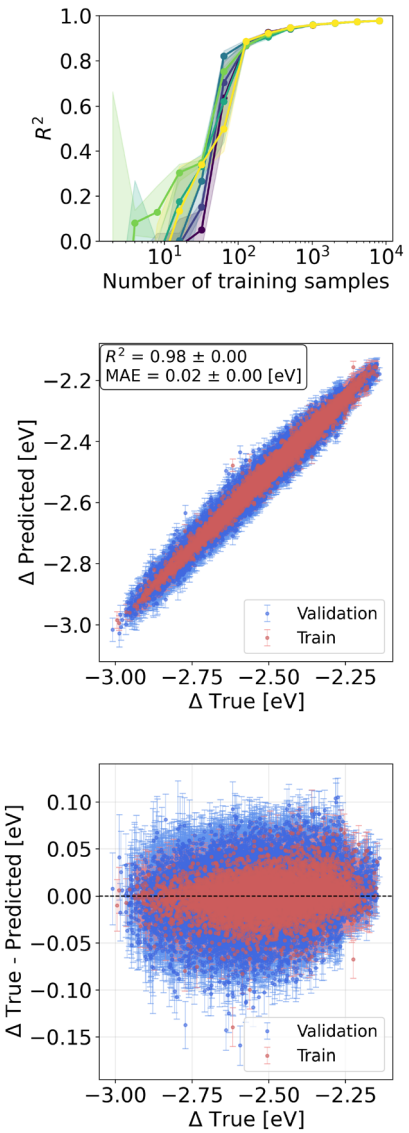

Fig. 6 Δ prediction from PDDF – top) learning curve observed when training a machine learning model to predict Pt nanoparticles stability from its PDDF representation. Middle) Corresponding parity plot between predicted and true activities for machine learning model trained with 8192 datapoints selected *via* an active learning protocol. Bottom) Corresponding error distribution of the difference true – predicted activities with respect to the true values.

3.3 Identifying most active stable catalyst

While the results show that GPR models trained on GCN or PDF representations achieve accurate predictions across a broad range of structures, (arguably) catalyst design is ultimately concerned with prioritizing the discovery of top-performing candidates. Thus, a key question remains whether this approach would enable to reliably identify most catalytically active and stable materials in a large pool of candidates.

To address this question, we discuss the outcome of a multi objective Bayesian optimization run starting from 2048 datapoints, aimed at identifying the Pareto front of Pt nanoparticle structures display the best activity–stability trade-offs. We note that, for practical purposes, in this case study we approximate catalyst stability in terms of its excess energy. To move beyond proof of concepts more realistic approaches to evaluate in-operando catalyst stability may be necessary.

Fig. 7 illustrates the outcome of the multi-objective Bayesian optimization (MOBO) procedure using the GCN- and PDDF-based structural representations. The full set of ~56 000 nanoparticle structures is shown as small transparent grey points, while the light-blue points denote the initial pool used to seed the optimization. The larger coloured markers represent the structures selected at successive MOBO iterations, with the colour scale indicating the iteration index. In both cases, the optimization rapidly identifies structures lying on a Pareto front (red line) overlapping with the true one (dashed grey line), demonstrating the efficiency of the surrogate models in navigating the Pt NPs activity–stability landscape for eORR. In Fig. 8, we report snapshots of the structures lying on the Pareto front, along with their number of atoms, size, and mass activity. Their visual inspection suggests the synthesis of nanoparticles with rather isotropic shapes, small (111) facets, and rich of convex sites between them, which indeed correspond to the most active sites for electrochemical ORR.

Interestingly, different exploration strategies emerge depending on the adopted representation. When using the PDDF descriptor, the optimization progressively samples structures with increasingly lower stability while maintaining high activity values. In contrast, the GCN-based optimization tends to explore structures with progressively decreasing activity along the Pareto front. These differences suggest that the two structural representations encode the activity–stability landscape in distinct ways (see also SI – Fig. S6 and

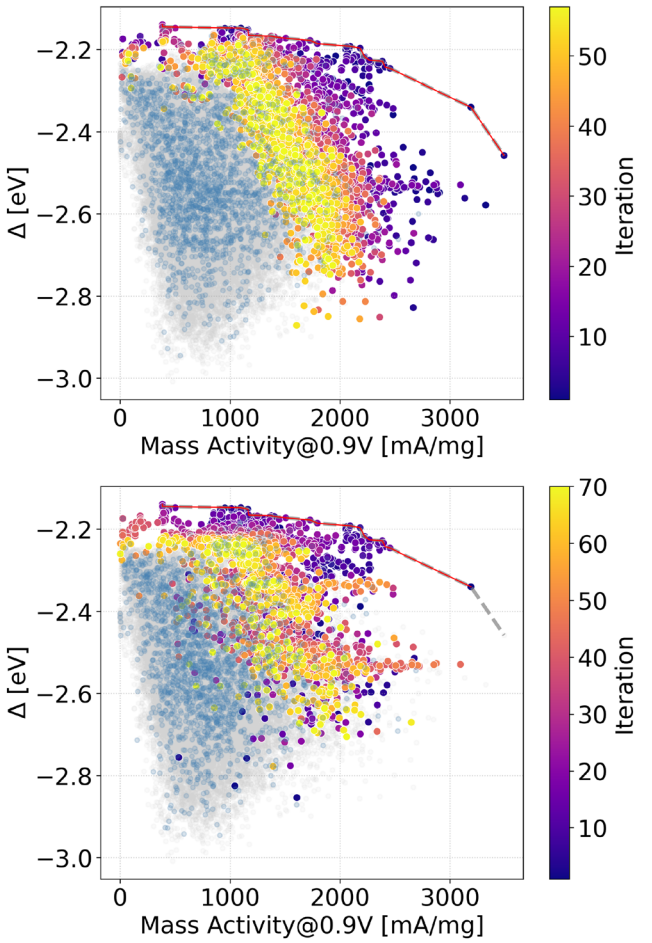

Fig. 7 Pareto front analysis of the multi-objective Bayesian optimization (MOBO), for the GCN (top), and the PDDF (bottom) NP representation. Gray points represent the full dataset of 52 658 structures. Steel blue points indicate the initial random training pool (2048 structures). Colored points show the structures sequentially selected by MOBO across 67 iterations, with color encoding the selection order from early (dark purple) to late (yellow) iterations. The red solid line marks the Pareto front of the MOBO-selected structures, while the dashed gray line shows the global Pareto front of the full dataset.

S7), ultimately guiding the Bayesian optimization toward different exploration trajectories despite converging to a similarly accurate approximation of the optimal Pareto frontier.

4 Conclusion

In conclusion, we presented a machine learning proof of concept to quantitatively chart structure–activity and structure–stability relationships in nanostructured catalyst. We do so by considering both structural representations accessible at a theoretical level, *i.e.*, generalized coordination number distribution, and experimentally, *i.e.*, pair distance distribution. Accurate machine learning predictions of structure–activity relationships were expected when representing nanoparticles using the distribution of atop generalized coordination numbers of their surface sites, also since these same were used as inputs in the microkinetic model that generated the ground truth activity data. For the case of a PDDF representation, the literature, *e.g.*, including but not limited to,[39,40] has shown that this quantity is of aid to qualitatively rationalizing catalytic trends. Through our machine learning approach, we show that PDDF measurements can be also used for quantitative prediction and identification of most active catalytic structures among a large pool of candidates.

Akin to efforts linking experimentally and theoretically derived quantities in heterogeneous catalysis,[42,43] we computationally validated the proposed framework using data generated from a microkinetic model that incorporates structure–property relationships previously validated against experiments. Importantly, however, the framework is not limited to microkinetic models. It can be readily adapted to work with data obtained from more computationally intensive simulations, such as kinetic Monte Carlo methods, and/or from (automated) experimental campaigns where high-throughput synthesis and characterization would enable iterative, on-the-fly model refinement.[44] This flexibility makes our approach a promising tool for accelerating catalyst discovery. In turn, future work could therefore extend the framework to more complex catalytic systems, including multicomponent materials, supported catalyst architectures, and more complex reaction networks.

Future implementations of the framework could also benefit from incorporating more sophisticated stability metrics derived from operando simulations or experimental characterization, which account for the full range of degradation pathways that may occur under operating conditions, including poisoning, oxidation, dissolution, reconstruction, and sintering.

Author contributions

K. R. conceived the project. F. B. provides funding and supervision. S. Z. and K. R. conducted the experiment(s). All authors analysed the results and reviewed the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

In relation to data availability we are glad to inform the interested reader that:

- .xyz coordinates of the Pt nanoparticles considered in the work, together with their GCN and PDDF are available at: https://doi.org/10.13130/RD$_$UNIMI/KWXQ58

- Python codes for GPR structure–activity prediction, active learning, and Bayesian optimization are available at: https://github.com/nanoMLMS/ML_GCN_PDDF_ORR

The Git repository also contains all the outputs of the codes when executed, including scripts to reproduce the graphs reported in this manuscript.

- Python codes for GCN and PDDF calculations are available at: https://github.com/nanoMLMS/pySNOW

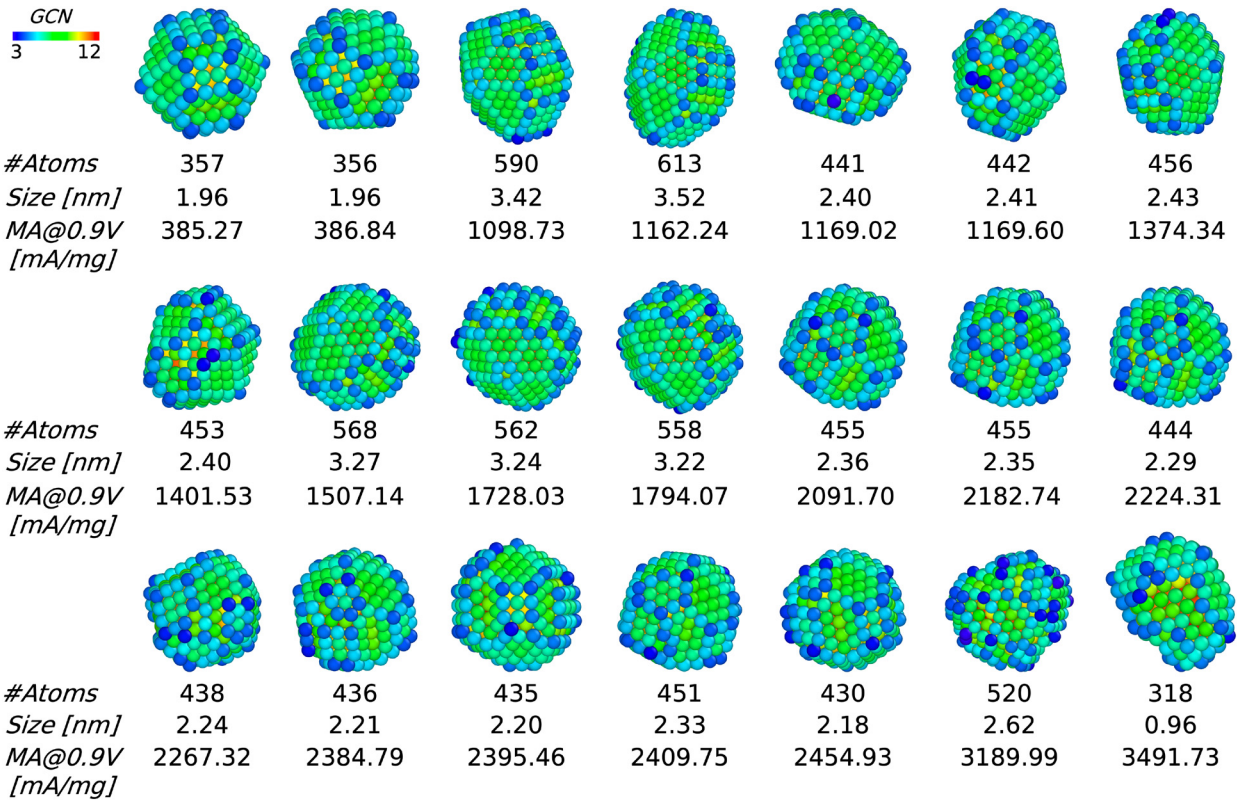

Fig. 8 Structural snapshots of the configurations located on the Pareto fronts in Fig. 7. All configurations are identified using both GCN and PDDF descriptors, except for the 318-atom structure, which is exclusive to the GCN feature representation. Atom colors correspond to GCN values, while the associated $MA_{@0.9V}$ and amount of atoms are reported for each structure.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6re00180g.

Acknowledgements

This article is based upon work from COST Action CA22154 – Data-driven Applications towards the Engineering of functional Materials: an Open Network (DAEMON) supported by COST (European Cooperation in Science and Technology). SZ acknowledges the University of Milan and ISC SRL for financially supporting her PhD studentship (D.M. 118/2023 PNRR). FB thanks the financial support by MONSTER, a project of the Italian National Centre for HPC, Big Data and Quantum computing (ICSC, CUP B93C22000620006) funded by the European Union – NextGenerationEU, through PNNR. All authors benefit of INDACO, a High-Performance Computing platform at the Università degli Studi di Milano, for computational resources. The work was also supported by the Research Foundation Flanders (FWO G076625N).

Notes and references

1 C. M. Friend and B. Xu, *Acc. Chem. Res.*, 2017, **50**, 517–521.
2 G. Rothenberg, *Catalysis: Concepts and Green Applications*, Wiley-VCH, Weinheim, Germany, 2008.
3 S. Mitchell, R. Qin, N. Zheng and J. Pérez-Ramírez, *Nat. Nanotechnol.*, 2021, **16**, 129–139.
4 H. Mistry, A. S. Varela, S. Kühl, P. Strasser and B. R. Cuenya, *Nat. Rev. Mater.*, 2016, **1**, year.
5 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, *Nat. Catal.*, 2022, **5**, 175–184.
6 L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski, R. Schlögl, A. Trunschke and M. Scheffler, *MRS Bull.*, 2021, **46**, 1016–1026.
7 S. Pablo-García, S. Morandi, R. A. Vargas-Hernández, K. Jorner, Ž. Ivković, N. López and A. Aspuru-Guzik, *Nat. Comput. Sci.*, 2023, **3**, 433–442.
8 J. Chen, X. Huang, C. Hua, Y. He and P. Schwaller, *Nat. Commun.*, 2025, **16**, 3232.
9 J. Lan, A. Palizhati, M. Shuaibi, B. M. Wood, B. Wander, A. Das, M. Uyttendaele, C. L. Zitnick and Z. W. Ulissi, *npj Comput. Mater.*, 2023, **9**, 172.
10 J. M. Mauß, K. S. Kley, R. Khobragade, N.-K. Tran, J. D. Bellis, F. Schüth, M. Scheffler and L. Foppa, *ACS Catal.*, 2025, **15**(15), 12652–12665.
11 L. Braunwarth, C. Jung and T. Jacob, *Top. Catal.*, 2020, **63**, 1647–1657.
12 V. Viswanathan and F. Y.-F. Wang, *Nanoscale*, 2012, **4**, 5110–5117.
13 M. Shao, A. Peles and K. Shoemaker, *Nano Lett.*, 2011, **11**, 3714–3719.
14 W. Sheng, S. Chen, E. Vescovo and Y. Shao-Horn, *J. Electrochem. Soc.*, 2011, **159**, B96–B103.

15 M. Nesselberger, S. Ashton, J. C. Meier, I. Katsounaros, K. J. J. Mayrhofer and M. Arenz, *J. Am. Chem. Soc.*, 2011, **133**, 17428–17433.

16 G. Tritsaris, J. Greeley, J. Rossmeisl and J. K. Nørskov, *Catal. Lett.*, 2011, **141**, 909–913.

17 D. Li, C. Wang, D. Strmcnik, D. Tripkovic, X. Sun, Y. Kang, M. Chi, J. Snyder, D. V. D. Vliet, Y.-C. Tsai, V. Stamenkovic, S. Sun and N. Markovic, *Energy Environ. Sci.*, 2014, **7**, 4061–4069.

18 B. Garlyyev, K. Kratzl, M. Rück, J. Michalička, J. Fichtner, J. M. Macak, T. Kratky, S. Günther, M. Cokoja, A. S. Bandarenka, A. Gagliardi and R. A. Fischer, *Angew. Chem., Int. Ed.*, 2019, **58**, 9596–9600.

19 F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet and D. Loffreda, *Angew. Chem., Int. Ed.*, 2014, **53**, 8316–8319.

20 K. Rossi, G. G. Asara and F. Baletto, *ACS Catal.*, 2020, **10**, 3520–3533.

21 J. Timoshenko, D. Lu, Y. Lin and A. I. Frenkel, *J. Phys. Chem. Lett.*, 2017, **8**, 5091–5098.

22 L. Delgado-Callico, K. Rossi, R. Pinto-Miles, P. Salzbrenner and F. Baletto, *Nanoscale*, 2021, **13**, 1172–1180.

23 J. I. Myung, J. R. Deneault, J. Chang, I. Kang, B. Maruyama and M. A. Pitt, *Digital Discovery*, 2025, **4**, 464–476.

24 S. Zinzani and F. Baletto, *Phys. Chem. Chem. Phys.*, 2024, **26**, 21965–21973.

25 S. Zinzani, M. Vanzan, R. M. Jones and F. Baletto, *Small Struct.*, 2026, **7**, e202500891.

26 F. Baletto and R. Ferrando, *Rev. Mod. Phys.*, 2005, **77**, 371–423.

27 F. Baletto, *J. Phys.: Condens. Matter*, 2019, **31**, 113001.

28 F. Calle-Vallejo, D. Loffreda, M. T. M. Koper and P. Sautet, *Nat. Chem.*, 2015, **7**, 403–410.

29 F. Calle-Vallejo, J. Tymoczko, V. Colic, Q. H. Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann and A. S. Bandarenka, *Science*, 2015, **350**, 185–189.

30 G. Becatti, S. Zinzani, D. Alimonti and F. Baletto, *pySNOW: a Python Suite for the NanO-World*, 2024, https://github.com/nanoMLMS/pySNOW.

31 F. Calle-Vallejo, M. D. Pohl, D. Reinisch, D. Loffreda, P. Sautet and A. S. Bandarenka, *Chem. Sci.*, 2017, **8**, 2283–2289.

32 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

33 N. Srinivas, A. Krause, S. M. Kakade and M. W. Seeger, *IEEE Trans. Inf. Theory*, 2012, **58**, 3250–3265.

34 J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger and A. G. Wilson, *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

35 B. Garlyyev, K. Kratzl, M. Rück, J. Michalička, J. Fichtner, J. M. Macak, T. Kratky, S. Günther, M. Cokoja, A. S. Bandarenka, A. Gagliardi and R. A. Fischer, *Angew. Chem., Int. Ed.*, 2019, **58**, 9596–9600.

36 M. Li, Z. Zhao, T. Cheng, A. Fortunelli, C.-Y. Chen, R. Yu, Q. Zhang, L. Gu, B. V. Merinov, Z. Lin, E. Zhu, T. Yu, Q. Jia, J. Guo, L. Zhang, W. A. Goddard III, Y. Huang and X. Duan, *Science*, 2016, **354**, 1414–1419.

37 M. W. Terban and S. J. L. Billinge, *Chem. Rev.*, 2022, **122**, 1208–1272.

38 E. Telari, A. Tinti, M. Settem, L. Maragliano, R. Ferrando and A. Giacomello, *ACS Nano*, 2023, **17**, 21287–21296.

39 R. Chattot, T. Asset, P. Bordet, J. Drnec, L. Dubau and F. Maillard, *ACS Catal.*, 2017, **7**, 398–408.

40 R. Chattot, P. Bordet, I. Martens, J. Drnec, L. Dubau and F. Maillard, *ACS Catal.*, 2020, **10**, 9046–9056.

41 C. Zeni, K. Rossi, A. Glielmo, Á. Fekete, N. Gaston, F. Baletto and A. D. Vita, *J. Chem. Phys.*, 2018, **148**, 241739.

42 X. Wang, S. Jiang, W. Hu, S. Ye, T. Wang, F. Wu, L. Yang, X. Li, G. Zhang, X. Chen, J. Jiang and Y. Luo, *J. Am. Chem. Soc.*, 2022, **144**, 16069–16076.

43 J. L. Lansford and D. G. Vlachos, *Nat. Commun.*, 2020, **11**, 1513.

44 A. S. Anker, J. H. Jensen, M. Gonzalez-Duque, R. Moreno, A. Smolska, M. Juelsholt, V. Hardion, M. R. V. Jorgensen, A. Faina, J. Quinson, K. Stoy and T. Vegge, *arXiv*, 2025, preprint, arXiv:2505.13571, DOI: 10.48550/arXiv.2505.13571.