Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation

Wenbin Xu\textsuperscript{1,2}, Karsten Reuter\textsuperscript{2} and Mie Andersen\textsuperscript{3,4 ,✉}

Computational screening in heterogeneous catalysis relies increasingly on machine learning models for predicting key input parameters due to the high cost of computing these directly using first-principles methods. This becomes especially relevant when considering complex materials spaces such as alloys, or complex reaction mechanisms with adsorbates that may exhibit bi- or higher-dentate adsorption motifs. Here we present a data-efficient approach to the prediction of binding motifs and associated adsorption enthalpies of complex adsorbates at transition metals and their alloys based on a customized Wasserstein Weisfeiler–Lehman graph kernel and Gaussian process regression. The model shows good predictive performance, not only for the elemental transition metals on which it was trained, but also for an alloy based on these transition metals. Furthermore, incorporation of minimal new training data allows for predicting an out-of-domain transition metal. We believe the model may be useful in active learning approaches, for which we present an ensemble uncertainty estimation approach.

Many surface catalytic reactions of vital importance to our society—such as Fischer–Tropsch, methanol or higher oxygenate syntheses—have complex reaction mechanisms with numerous intermediates ranging from atoms and simple molecules to (possibly oxygenated) C1, C2 or larger fragments. It is well-known that modeling of these latter complex species at transition metal catalysts must account for their ability to exhibit a wide range of adsorption motifs, including mono-, bi- and higher-dentate adsorption modes\textsuperscript{1–3}. Density-functional theory (DFT) with van der Waals corrections can, in principle, provide the energetics of such adsorption motifs at moderate cost and satisfactory accuracy\textsuperscript{4}. Nevertheless, the identification of the most stable adsorption motifs of adsorbates involved in ethanol synthesis on a simple monometallic catalyst such as Rh(111) is already a formidable task\textsuperscript{5–8}, and the investigation of broader classes of materials such as transition metal alloys is generally out of reach due to the combinatorial explosion of possible active sites and adsorption motifs.

Machine learning models have already shown their potential for replacing expensive DFT calculations to tackle the screening of large materials spaces for accelerated catalyst discovery\textsuperscript{9–13}. However, most works so far have been limited in scope to the consideration of atoms or small molecules with monodentate adsorption motifs. For these simple species, models now routinely achieve the prediction of adsorption enthalpies with a root-mean-square error (r.m.s.e.) of around 0.1–0.2 eV, which is then comparable with the intrinsic DFT accuracy\textsuperscript{11,13–15}. Unfortunately, most of these methods cannot easily be extended to complex adsorbates with bi- or higher-dentate adsorption motifs. One notable attempt to treat complex adsorbates is provided in the Open Catalyst Project where the direct prediction of relaxed adsorption enthalpies of complex adsorbates at transition metals and their alloys based on a customized Wasserstein Weisfeiler–Lehman graph kernel and Gaussian process regression. The model shows good predictive performance, not only for the elemental transition metals on which it was trained, but also for an alloy based on these transition metals. Furthermore, incorporation of minimal new training data allows for predicting an out-of-domain transition metal. We believe the model may be useful in active learning approaches, for which we present an ensemble uncertainty estimation approach.

\textsuperscript{1}Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Garching, Germany. \textsuperscript{2}Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany. \textsuperscript{3}Aarhus Institute of Advanced Studies, Aarhus University, Aarhus, Denmark. \textsuperscript{4}Department of Physics and Astronomy—Center for Interstellar Catalysis, Aarhus University, Aarhus, Denmark. \textsuperscript{✉}e-mail: mie@phys.au.dk

\textsuperscript{1}Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Garching, Germany. \textsuperscript{2}Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany. \textsuperscript{3}Aarhus Institute of Advanced Studies, Aarhus University, Aarhus, Denmark. \textsuperscript{4}Department of Physics and Astronomy—Center for Interstellar Catalysis, Aarhus University, Aarhus, Denmark. \textsuperscript{✉}e-mail: mie@phys.au.dk
enthalpies with a r.m.s.e. of about 0.2 eV and also shows good extrapolative performance for two test cases; bimetallic alloys made from elements present in our training data and out-of-domain elements, the latter however only after incorporation of adsorption enthalpies of atomic species on the new element into the training database. Finally, we show that data points with large prediction errors can be quite reliably captured from an ensemble uncertainty estimation approach.

**Results**

**WWL-GPR model.** The machine learning task in our work is to directly predict the relaxed adsorption enthalpies corresponding to a range of possible adsorption motifs represented as graphs. Thereby, for a given surface–adsorbate combination of interest, we obtain a spectrum of possible adsorption energies ranging from the most stable to metastable adsorption motifs. Microkinetic models used in catalyst screening often employ only the most stable adsorption energy obtained as input, however, distinct adsorption sites with less favorable adsorption energies could be included as well. Our task is thus quite similar to the task denoted as initial state to relaxed energy (IS2RE) in the Open Catalyst Project, however, we do not directly use the initial state geometry, but only its graph representation. We note that an entirely different approach to this task is to train a machine learning interatomic potential to relax the initial structure and thereby predict both the relaxed structure and adsorption enthalpy; however, such approaches are not a topic of this work.

Figure 1a depicts a schematic of our physics-inspired WWL-GPR model. We rely on graph representation, which is a versatile method for representing isolated molecules or crystal structures or the combined surface–adsorbate system in which every atom in the structure is a node with edges representing chemical bonds to neighboring atoms. Graph representation can be used in connection with neural networks, which generally requires very large training databases. As we are here interested in developing a data-efficient method, we focus on a kernel-based method (GPR) in connection with a customized version of the recently developed WWL graph kernel.

**Prediction of simple adsorbates.** We begin by evaluating the performance of the machine learning models for predicting a database of simple adsorbates with monodentate adsorption motifs (see Methods). We perform fivefold cross-validation, that is, the database is shuffled and partitioned into five equal-sized subsamples stratified by adsorbates. The training is then performed based on four of the subsamples while retaining the fifth subsample for validation. This is repeated five times until all data points have been used once for validation. Figure 2 shows the resulting parity plot of DFT-calculated versus machine learning-predicted adsorption enthalpies for the combined validation set from the five folds as well as violin plots of the absolute error distributions for SISSO and the GPR models. It should be noted here that the SISSO results are obtained using similar hyperparameters to those from our previous work (eight-dimensional rung three descriptor). In principle, we would expect a better performance for even more complex models than the r.m.s.e. value of 0.24 eV presented here; see Supplementary Fig. 5. However, the identification of more complex models is
computationally intractable with the SISSO method. Rather than raw performance, the merit of the SISSO approach is that the identified descriptors are (somewhat simple) analytical functions of the features, which are thus easier to interpret than black-box machine learning models. We also note that the reason for the different performance of the descriptors identified in the present work compared with our previous work is that here we train a single model on the entire database (single-task learning) to be able to make a direct comparison to the GPR models, whereas in our previous work separate fitting coefficients were used for each adsorbate (multitask learning). More information about the identified SISSO descriptors is provided in Supplementary Section 3.2.

The model complexity can be more easily tailored for the GPR models, and, after optimization of the relevant hyperparameters (see Supplementary Section 3.5 and Supplementary Table 9), we obtain a r.m.s.e. of 0.13 eV independent of whether we use vector input (RBF-GPR) or graph representation (WWL-GPR). The maximum absolute error (maxAE) also decreases from 1.11 eV (SISSO) to around 0.60 eV in the GPR models. Based on the similar performance of the two GPR models, we can conclude that there is no added value from employing graph representation for the simple adsorbates. The reactivity is apparently already captured by the averaged surface atom features and the adsorbate-specific features used in the RBF-GPR model.

Finally, the XGBoost method represents an ensemble-based machine learning method based on decision trees and gradient boosting, where trees are added one at a time to improve on the residuals of the previous model [33]. Here we find that it performs similarly to the GPR models (see Supplementary Fig. 8), with an r.m.s.e. of 0.12 eV. On the basis of this similar performance of state-of-the-art methods, we believe that we are at the limit of the machine learning accuracy achievable for simple adsorbates with the available dataset and feature representation.

**Prediction of complex adsorbates.** We next turn to a database of complex adsorbates with 41 different adsorbates in mono-, bi- and higher-dentate adsorption motifs on surfaces of Cu, Co, Pd and Rh (see Methods). As we already concluded in the preceding section on simple adsorbates that single-task SISSO is not competitive in terms of performance, we focus here only on the GPR models and XGBoost. The fivefold cross-validation results presented in Fig. 3 show that the graph-based WWL-GPR model has a superior performance (r.m.s.e. of 0.18 eV) for this more challenging database compared with RBF-GPR (r.m.s.e. of 0.47 eV). The maxAE also decreases from 2.23 eV (RBF-GPR) to 0.92 eV in the WWL-GPR model. The XGBoost method clearly outperforms RBF-GPR, with a r.m.s.e. of 0.23 eV, which is possibly related to the advantages of its ensemble-based approach; however, it is still inferior to WWL-GPR. We attribute this to the importance of the graph representation for complex adsorbates, which is present in the WWL-GPR model but missing in the vector-based models. A learning curve for the WWL-GPR model is presented in Supplementary Fig. 6, which shows that an r.m.s.e. of 0.3 eV can be achieved by only training on 30% of the database (~500 data points) and a r.m.s.e. of 0.2 eV is achieved at 70% of the database (~1,200 data points). A visualization of the prediction accuracy for adsorption motifs of one selected adsorbate (CHCO) on one selected surface (Cu(211)) is given in Supplementary Fig. 7.

To visualize what trends the WWL-GPR model has identified in the complex adsorbates database, we present in Fig. 4 a kernel principal component analysis, which is a non-linear dimensionality reduction technique. Specifically, we here present the two dimensions that explain the largest fraction of the variance. Points that are close together in this space are similar in the feature space. The analysis of the entire complex adsorbate database in Fig. 4a shows that the different metals are distinguished as parallel clusters, where for each cluster there is a similar distribution of subclusters containing the individual adsorbates. In Fig. 4b the same analysis is presented for only one metal (here, Rh, but similar results are obtained for the other transition metals). Again, the different adsorbates form clusters, where each point in a cluster corresponds to a separate adsorption motif of the adsorbate. A similar clustering cannot be observed in KPCA plots for the RBF-GPR model (see Supplementary Fig. 9), which is probably related to the fact that this model does not have any structural information on the different adsorbates and their associated adsorption motifs due to the lack of graph representation.

Having established the excellent interpolation performance of the WWL-GPR model, we next assess the predictive performance of the model for extrapolation tasks concerning data that are dissimilar to those in the training database, that is, an out-of-domain prediction. This is highly important for the practical application of the model to catalyst screening. The two tasks we consider are: (1) predictions for a bimetallic catalyst, that is, an alloy of elemental metals present in our database; and (2) predictions for an out-of-domain element when merely incorporating the adsorption enthalpies of atomic species (C, H, and O) at the new element into the database. For these tasks we selected eight adsorbates comprising atomic species and larger molecules, and including some with bidentate adsorption motifs (see Supplementary Table 4).

As it has previously been emphasized in the literature that a careful choice of regularization can substantially improve the robustness of a model in extrapolative, data-poor regimes [33], we reoptimized the hyperparameters for the extrapolation tasks. Specifically, we...
where r.m.s.e.interpolation is the r.m.s.e. for the original complex adsorbates for Pt). As the aim is to find hyperparameters well suited for extrapolation, this task was given a higher weight (two) in the loss function than that of the interpolation task (one). Comparing the hyperparameters obtained previously for the complex adsorbates database (including atomic adsorption enthalpies for Pt) and r.m.s.e.extrapolation is the r.m.s.e. of the dataset for CuCo and Pt (only the complex adsorbates for Pt). As the aim is to find hyperparameters well suited for extrapolation, this task was given a higher weight (two) in the loss function than that of the interpolation task (one). Comparing the hyperparameters obtained previously for the complex adsorbates database (see the base case in Supplementary Table 10) with the new hyperparameters optimized specifically for the extrapolation tasks (see the base case in Supplementary Table 11), we see that indeed both the length scale and the regularization term increase for the extrapolation tasks, resulting in a smoother machine learning model, which is consistent with the previous literature observations. The r.m.s.e. values obtained for the new hyperparameters are 0.25 eV for interpolation within the complex adsorbate database, 0.23 eV for the CuCo alloy and 0.30 eV for Pt. Finally, we carry out a true extrapolation test to assess whether the new hyperparameters would also be accurate for yet another bimetallic alloy (here we chose PdRh) and yet another out-of-domain element (here we chose Ru); see Supplementary Table 4. We are able to obtain a very good extrapolation performance with a r.m.s.e. of 0.23 eV for PdRh and 0.23 eV for Ru. We note here that Ru is apparently easier to extrapolate to than Pt (on the basis of the lower r.m.s.e. obtained), which signifies that it must somehow be more similar to the elements present in the complex alloys database. Our results also show that an out-of-domain element is generally harder to predict than an alloy of known elements, even when incorporating some minimal information about the unknown element into the training database through the atomic adsorption enthalpies. We would expect the performance for Pt to improve if more adsorbates were added to the training database.

Uncertainty quantification. Up till now we have demonstrated that our WWL-GPR model can be applied, with r.m.s.e. values of around 0.2–0.3 eV, to flat and stepped metal and bimetallic catalysts, as long as some (at least minimal) training data involving the considered elemental metals are provided; however, apart from knowing which average r.m.s.e. value to expect, it is also useful to be able to directly assess the expected uncertainty on a single predicted data point. For example, uncertainty quantification combined with sensitivity analysis of microkinetic models can be used to assess error propagation and the extent to which conclusions drawn from a model are robust to input parameter uncertainty. Furthermore, uncertainty quantification is used in active learning approaches, where the training database is iteratively updated through selected DFT calculations, for example, of data points with a high estimated uncertainty.
For calibration, sharpness and dispersion value (small error estimates) and a large dispersion value (large error estimates), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution). A useful uncertainty quantification method should have a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution), a small miscalibration area (that is, a good match between the expected and observed cumulative error distribution). However, with our primary interest being active learning, it is much more intriguing to see that the ensemble model does a better job than the single model at assigning a high uncertainty to data points with a high actual prediction error. In particular, the group of points with an estimated uncertainty higher than 0.2 eV in the ensemble model includes the largest prediction errors, whereas this is not the case for the single model, which actually assigns quite a low uncertainty to some of the largest prediction errors. We therefore conclude that the ensemble model is best suited for active learning approaches. We note here that we do not expect any quantitative match between the absolute error and the uncertainty in Fig. 5, partly as these are not directly comparable quantities (one is a standard deviation and the other an absolute error), and partly because it has been shown that specific calibration measures\(^4^3\) are required for quantitatively accurate uncertainty quantification in both single and ensemble GPR models.

**Discussion**

We begin by discussing the origin of the superior performance of the WWL-GPR model over the vector-based RBF-GPR and XGBoost models. First of all, we note that it is not surprising that for complex adsorbates, simply accounting for the surface and adsorbate in terms of features averaged over the atoms directly involved in the bonding, as well as the global features of the adsorbate (for example, HOMO/LUMO levels) and clean surface (for example work function), as done in the vector-based models, is insufficient. By contrast, the graph representation provides direct access to structural information on the system, that is, the number and types of atoms in the adsorbate and how these atoms connect to each other and to the surface, possibly in complex bi- or higher-dentate adsorption motifs. Atom-specific features related to the local electronic or geometric structure can be directly used as node attributes (for...
example, through SOAP descriptors), and we can introduce surface adsorption-motivated hyperparameters as discussed above and illustrated in Fig. 1c. The main remaining limitation of our model is that it cannot be expected to handle cases where the adsorbate dissociates or the surface reconstructs upon the adsorption event, as it—by contrast to machine learning force fields—does not predict the entire potential energy surface of the system, but only discrete minima corresponding to the adsorbed states. Furthermore, it relies on user-specified features, which would have to be adjusted for the consideration of other materials classes, for example, metal oxides.

Based on the demonstrated extrapolation performance, we trust that our WWL-GPR model could be useful for catalyst screening purposes, for example, for exploring reactions with complex adsorbates on alloy surfaces. Here the complexity encountered from the many possible adsorption motifs of each adsorbate on each type of alloy surface makes direct DFT investigations computationally intractable, while reliable machine learning force fields or density-functional tight-binding methods for the simultaneous treatment of many different adsorbates and/or alloy surfaces are still difficult to obtain.

We envision that it could be particularly interesting to apply our model in the context of an active learning strategy, where the training database is iteratively expanded towards catalytically interesting and/or previously poorly explored regions of the catalyst space. Key advantages of our data-efficient GPR model in this regard are the low training cost (compared to, for example, deep neural networks) and the demonstrated uncertainty quantification. For active learning purposes, we also recommend using the model with different hyperparameter settings depending on the exploitative or learning purposes, we also recommend using the model with different hyperparameter settings depending on the exploitative or explorative nature of the task at hand. Specifically, we can confirm the findings from past literature reports that hyperparameters characterized, among others, by larger length scale and regularization terms are beneficial for accurate predictions in data-poor regions of the catalyst space.

Methods

DFT databases. The machine learning models are trained and tested on two different databases, termed simple adsorbates and complex adsorbates. The former is taken from refs. 11,23. After a post-processing step, the database contains 1,422 different databases, termed simple adsorbates and complex adsorbates. The considered surfaces include the fcc(100), fcc(110), fcc(111) and fcc(211) facets of pure Ni, Cu, Ru, Pt, Pd, Ag, Ir, Pt and Au, the body-centered cubic (bcc) (210) facet of Fe as well as the stepped hexagonal close-packed (hcp) (0001) facet of Co. For all catalysts, the database contains the adsorbates on the four single-atom alloys Ag@Cu, Pt@Rh, Pd@Ir and Au@Ni (that is, single-atom Ag, Pt, Pd or Au dispersed in the surface of another host metal), and the four AB bimetallic alloys AgPd, IrRu, PtRh and AgAu.

The complex adsorbates database contains 1,679 data points and includes the adsorption enthalpies of eight simple adsorbates with monodentate coordination: C, H, O, CO, CH, CH2 and CH3. The considered surfaces include the fcc(100), fcc(110), fcc(111) and fcc(211) facets of pure Ni, Cu, Ru, Pt, Pd, Ag, Ir, Pt and Au, the body-centered cubic (bcc) (210) facet of Fe as well as the stepped hexagonal close-packed (hcp) (0001) facet of Co. For all catalysts, the database contains the adsorbates on the four single-atom alloys Ag@Cu, Pt@Rh, Pd@Ir and Au@Ni (that is, single-atom Ag, Pt, Pd or Au dispersed in the surface of another host metal), and the four AB bimetallic alloys AgPd, IrRu, PtRh and AgAu.

The complex adsorbates database contains 1,679 data points and includes 41 different small and large adsorbates involved in ethanol synthesis on fcc(111) and fcc(211) facets of Cu, Rh, Pd and Co. Examples of complex adsorbates are CHCO, CCHOH, CH2OH, O2, OH and CH3OH; the full list of adsorbates is provided in Supplementary Table 3. Furthermore, selected adsorbates are calculated at the CoCo(111), PdRh(111), Pt(111), Ru(111), Pt(211) and Ru(211) surfaces for model testing purposes. The adsorbates contain up to nine atoms and cover mono-, bi- and higher-dentate adsorption modes. The database is constructed using an automated workflow and DFT settings that are compatible with the simple adsorbates database (Quantum Espresso code14 and the Bayesian error estimation functional with van der Waals correlation (BEET-vdW) functional45). Further computational details and overviews of both databases are given below and in Supplementary Section 1.

Database construction and workflow. The initial geometries of the surface–adsorbate systems are generated using the CatKit software. CatKit employs a graph representation of the surface atoms to enumerate mono- and bidentate adsorption sites, where the latter are defined by a neighboring node–edge pair of the graph. For each adsorbate, a manual tagging of the bonding atoms for mono- and bidentate adsorption motifs is required (see Supplementary Table 3). CatKit then adds the adsorbates at the enumerated adsorption sites by employing some simple geometric procedures to produce good estimates for the angles and bond lengths in the system. We note that CatKit does not generate all possible adsorption motifs (which would be computationally intractable), but only those that are judged as most plausible. This adds a human bias into the generation of the database. Furthermore, not all initial geometries generated are actually stable, but could transform into other structures during the DFT relaxation.

To overcome some of these limitations, we added the following steps to our computational workflow. During the DFT relaxation, we monitor the graph representation of the system and assign to it the following four cases: (1) if the graph representation is unchanged, the data point is simply added to our database (32.3% of cases); (2) if the structure transforms into another graph that is already enumerated in the CatKit database (15.9% of cases), then the calculation with the most favorable adsorption enthalpy is added to the database to avoid duplicates; (3) the calculation is discarded if the structure transforms into a non-valid graph, that is, a graph that is incompatible with our direct graph-based machine learning model (for example, adsorbate dissociation, surface reconstruction) (23.4% of cases); (4) if the structure transforms into a valid graph that was not enumerated by CatKit (8.8% of cases), then the database is updated with the initial graph representation and the new adsorption motif is added also for the other surfaces of interest. The last case (4), as well as large adsorbates whose initial adsorption motifs cannot be well controlled by CatKit, are the source of all higher-dentate adsorption motifs in our database (see Supplementary Fig. 3 for examples). Our workflow is implemented with AIDA28, which is a scalable computational infrastructure providing advanced automation to allow interfacing with external simulation software. In our case this entails customized python scripts interfacing with CatKit, the Atomic Simulation Environment software29 and the Quantum Espresso DFT code.

DFT computational details. For the DFT calculations of the complex adsorbates database, the calculations involving Co. The cutoff energy was set to 500 eV and 5,000 eV for the orbitals and charge density, respectively, and a Fermi-level smearing of 0.1 eV was used. The resulting adsorption enthalpies are formation energies referenced to gaseous CH, OH, CO and H2.

The features that require DFT calculations were obtained as follows. For the clean surfaces involved in both the simple and complex adsorbate databases, we first performed a geometry relaxation as outlined above. The projected density of states (PDOS) was calculated using the smeared-free tetrahedron method and an energy spacing of 0.01 eV. We used a (14×14) k-point grid for the pure metal fcc and bcc slabs, a (7×21) grid for the S22425 fcc alloys, a (14×21) for the S22421 fcc alloys and a (7×7) grid for the CuCo(111) and PdRh(111) alloy surfaces involved in the extrapolation tasks. For the calculation of band moments, we integrated empty bands up to the energy above the Fermi level where the PDOS had fallen below a value of 0.01 Å−1 eV−1. The features involving the density of states at the Fermi level were calculated using a smearing of 0.1 eV in the PDOS calculation, and the PDOS was averaged over the interval ±0.1 eV around the Fermi level. For the calculation of adsorbate-specific features, we carried out a structural optimization of the isolated adsorbate positioned in a cubic supercell with a side length of 15 Å. We used a Fermi-level smearing of 0.01 eV and the Brillouin zone integration was performed using the Gamma point only.

Further details on machine learning models. The WWL-GPR model is compared with three other machine learning models (Sissso, RBG-FFP and XGBoost) that do not use graph representation, but input in vector form with features of the clean surface and the isolated adsorbates. The features used in the vector-based models are specific to the adsorbate. For adsorbate-specific features, we calculated averaging over the metal atoms to which the adsorbate coordinates (clean surface features) or the bonding atoms of the adsorbate (isolated adsorbate features). The WWL-GPR model also uses atom-specific features as node attributes, for example, electronic properties of individual surface atoms or features of the local geometry of the clean surface and isolated adsorbate through SOAP (Overlap of Atomic Positions) descriptors. All of the details on the features used in the compared machine learning models are provided in Supplementary Section 2.
Supplementary Section 3 provides more information about each of the models, including a more in-depth discussion on the hyperparameters. When comparing the RBF-GPR and WWL-GPR models, it is interesting to note that although the WWL-GPR model finds that the optimal value is one node distance for both inner and outer cutoff for the simple adsorbates database (that is, mostly the atoms directly involved in surface–adsorbate bonding are judged important), the optimal inner and outer cutoffs (weights) are one (0.60) and two (0.06) node distances, respectively, for the complex adsorbates database, see Supplementary Table 9 (that is, also atoms neighboring the immediately bonding atoms are judged important, although with smaller weights). The effect of more distant atoms is not taken into account in the vector-based models, which then possibly relates to their decreased performance for complex adsorbates. Note also that during the node embedding scheme of the WWL graph kernel, the node attribute of every atom is updated with information about the node attributes of the neighboring atoms. In Supplementary Section 3.3.1.2. That is, even if weights beyond the outer cutoff are zero, the atoms there can still have a non-negligible influence on the kernel value.

It should be emphasized that the WWL-GPR model leverages only features from the initial guess geometry, specifically, the graph connectivity, and electronic and geometric features calculated from the clean surface and isolated adsorbate. From a computational screening point of view this is essential for keeping the computational cost of model predictions low. The computationally most intensive part of the model prediction is the DFT calculation of the clean surface to obtain the node attributes (for example bond-moments) for the surface atoms. However, given that we target 41 different adsorbates in various possible adsorption motifs for each surface, this is still a low-cost-per-machine learning prediction.

For SISSO, we previously used an approach to target simple adsorbates where the free parameters of the identified models were fitted to each adsorbate separately.""
39. Döpking, S. & Matera, S. Error propagation in first-principles kinetic monte carlo simulation. *Chem. Phys. Lett.* **674**, 28–32 (2017).

40. Flores, R. A. et al. Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction. *Chem. Mater.* **32**, 5854–5863 (2020).

41. Kunkel, C., Margraf, J. T., Chen, K., Oberhofer, H. & Reuter, K. Active discovery of organic semiconductors. *Nat. Commun.* **12**, 1–11 (2021).

42. Tran, K. et al. Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn. Sci. Technol.* **1**, 025006 (2020).

43. Palmer, G. et al. Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Comput. Mater.* **8**, 1–9 (2022).

44. Kuleshov, V., Fenner, N. & Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning* 2796–2804 (MLR Press, 2018).

45. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).

46. Wellendorff, J. et al. Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).

47. Huber, S. P. et al. AiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7**, 300 (2020).

48. Xu, W., Reuter, K. & Andersen, M. Predicting Binding Motifs of Complex Adsorbates Using Machine Learning with a Physics-Inspired Graph Representation (Zenodo, 2022); https://doi.org/10.5281/zenodo.6640198

49. Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).

50. Dal Corso, A. Pseudopotentials periodic table: from H to Pu. *Comput. Mater. Sci.* **95**, 337–350 (2014).

51. Garrison, K. F., Bennett, J. W., Rabe, K. M. & Vanderbilt, D. Pseudopotentials for high-throughput DFT calculations. *Comput. Mater. Sci.* **81**, 446–452 (2014).

52. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

53. Esterházi, J. A., Goldsmith, B. R. & Linic, S. Theory-guided machine learning finds geometric structure–property relationships for chemisorption on subsurface alloys. *Chem* **6**, 3100–3117 (2020).

54. Andersen, M. & Reuter, K. Adsorption enthalpies for catalysis modeling through machine-learned descriptors. *Acc. Chem. Res.* **54**, 2741–2749 (2021).

55. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).