Chemical Science

EDGE ARTICLE

Check for updates

Cite this: *Chem. Sci.*, 2018, **9**, 7069

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

Machine learning meets volcano plots: computational discovery of cross-coupling catalysts†

Benjamin Meyer,[ac] Boodsarin Sawatlon,[ac] Stefan Heinen,[bc] O. Anatole von Lilienfeld*[bc] and Clémence Corminboeuf*[ac]

Received 29th April 2018
Accepted 12th July 2018

DOI: 10.1039/c8sc01949e

rsc.li/chemical-science

The application of modern machine learning to challenges in atomistic simulation is gaining attraction. We present new machine learning models that can predict the energy of the oxidative addition process between a transition metal complex and a substrate for C–C cross-coupling reactions. In turn, this quantity can be used as a descriptor to estimate the activity of homogeneous catalysts using molecular volcano plots. The versatility of this approach is illustrated for vast libraries of organometallic catalysts based on Pt, Pd, Ni, Cu, Ag, and Au combined with 91 ligands. Out-of-sample machine learning predictions were made on a total of 18 062 compounds leading to 557 catalyst candidates falling into the ideal thermodynamic window. This number was further refined by searching for candidates with an estimated price lower than 10 US$ per mmol. The 37 catalyst finalists are dominated by palladium phosphine ligand combinations but also include the earth abundant transition metal (Cu) with less common ligands. Our results indicate that modern statistical learning techniques can be applied to the computational discovery of readily available and promising catalyst candidates.

1 Introduction

Chemists constantly pursue new molecular systems that provide increasingly higher yields and better control of selectivity. Rather than blindly searching for promising catalysts to meet their needs, numerous tools that aid in identifying the most appropriate species have been developed. These range from high-throughput screening[1,2] (including combinatorial methods[3,4]), which quickly evaluates reaction conditions and the structures of catalysts, to multidimensional modeling based on a design of experiments (DoE),[5] that relates steric and structural descriptors (*e.g.*, Charton values and Sterimol parameters) to enantioselectivity. Such methods have found broad application in asymmetric homogeneous catalysis.[6–14] On the other hand, the tremendous increase in computer power accompanied by methodological advancements has also made computational studies of catalytic processes commonplace.[15] While virtually any catalytic system can be subjected to computational analysis, often the conclusions reached are not transferable and provide little insight into the best way to develop more active and selective catalysts. Thus, a tool that assesses the properties of untested catalysts based on a simple energetic or structural criterion would rapidly accelerate the discovery pace of new species. Indeed, similar concepts involving the mapping of a difficult to determine quantity onto an easily obtained variable have been a central pillar of catalysis and physical organic chemistry for more than 80 years, and are at the core of familiar concepts such as the Bell–Evans–Polanyi principle,[16,17] the Hammett equation,[18–21] or the Brønsted catalysis equation.[22] Today, volcano plots,[23,24] which relate easily accessible descriptor variables directly to catalytic performance, accomplish this objective and find regular use in the fields of heterogeneous catalysis[25–27] and electrocatalysis.[28–34]

Based on knowledge of a chosen descriptor variable, volcano plots function by discriminating catalytic performance using Sabatier's principle.[35] Sabatier conceived the notion of an ideal catalyst that should not bind a substrate too strongly or too weakly. The unique volcano shape facilitates rapid discrimination of catalytic activity. Species positioned highest on the plot (generally on or near the volcano plateau or peak) have the best profiles and fulfill Sabatier's principle. Species located along the left- and right-slopes have less ideal profiles and can be characterized as having either overly strong (left) or overly weak (right) substrate/catalyst interactions. While being commonly

[a] *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. E-mail: clemence.corminboeuf@epfl.ch*

[b] *Institute of Physical Chemistry, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland. E-mail: anatole.vonlilienfeld@unibas.ch*

[c] *National Center for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

† Electronic supplementary information (ESI) available: Details of the ligand dataset, machine learning predictions, price of metals and ligands and all Cartesian coordinates (.xyz) are included in the file SuppInfo.tar.bz2. See DOI: 10.1039/c8sc01949e

used in heterogeneous and electrocatalysis, and frequently invoked for homogeneous systems,[36–38] only recently have these appealing tools been concretely realized for molecular catalysts.[39] Corminboeuf and co-workers have pioneered the use of molecular volcano plots to study various aspects of prototypical C–C cross-coupling reactions in order to gauge the feasibility of using these tools to identify attractive homogeneous catalysts.[39–42] Subsequent work has also focused on adapting volcano plots for applications in homogeneous catalysis *via* the inclusion of kinetics (as opposed to the typically used thermodynamic based criteria) of the catalytic cycle.[43]

The use of molecular volcano plots involves establishing linear scaling relationships that relate the quantitative value of a descriptor to the thermodynamic or kinetic performance of the catalyst. As such, this tool has clear utility in high-throughput screening applications that search for prospective catalysts by computing the value of this descriptor for any species desired. However, currently both the geometries and energies associated with multiple forms of each catalysts must be determined through a relatively slow process involving density functional theory computations. Clearly, increasing the speed at which the descriptor variable can be determined would result in an overall increase in the discovery pace of new catalysts because prospective species could be screened more rapidly. One route with the potential to provide virtually instantaneous access to the descriptor involves the application of quantum machine learning (ML) models, *i.e.*, ML models which can be trained on, and used to predict, quantum properties.[44–46] The application of ML models to estimate volcano plot energy descriptors offers increased speed for two reasons: first, the energy based value can be immediately accessed for any desired species, and second, the need to establish a precise geometry of the catalyst can be circumvented by also including this task into the ML model, as already demonstrated within the Δ-ML approach.[47] As such, the ML model can predict an accurate descriptor value from an approximated 3D structure of a catalyst.

While, generally speaking, applications of machine learning methods in chemistry are still in their infancy, their use has begun to appear in the fields of materials science[48–53] and catalysis.[54–61] For example, a gradient-boosting regression method[62] has been used to predict the d-band center of mono and bimetallic surfaces[63] and to estimate CO adsorption energies on Pt nanoparticles,[64] while a local similarity kernel could predict the catalytic activity of nanoparticles.[65] Moreover, applications of support vector machines (SVMs)[66] were able to anticipate CO_2 uptake in metal organic frameworks (MOFs)[67] by developing an atomic property-weighted radial distribution function (AP-RDF) based descriptor[68] that captures geometric and chemical features of periodic systems. Predictive structure–reactivity models have identified promising Pt-based electrocatalysts for the oxygen reduction reaction,[69] while artificial neural networks (ANNs) have recognized multimetallic alloys possessing high selectivity for electrochemical CO_2 reduction to C_2 species.[70,71] Recently, Nørskov investigated various machine learning based approaches[72] to systematically search for the active sites of bimetallic (nickel gallium) nanoparticles,[73] to construct Pourbaix surface phase diagrams,[74] and to identify probable mechanisms of hydrocarbon–syngas reactions on rhodium(111).[75] Rappe and co-workers also exploited the regularized random forest machine learning algorithm,[76] and discovered the key role played by structure and charge descriptors (namely the Ni–Ni bond length and the Ni residual charge) in the hydrogen evolution reaction activity of $Ni_2P(0001)$.

Despite the considerable amount of progress in applying ML models to chemical problems, each of the aforementioned contributions tackled issues surrounding heterogeneous catalysis, while ML applications to homogeneous catalysis remain exceedingly rare.[54,77] Significant advances with ML models to obtain fundamental molecular electronic properties (*e.g.*, atomization or total energies of molecules) have been made,[78–85] however, the prediction of catalytic cycle intermediates energies has never been attempted to the best of our knowledge. The purpose of this work is to demonstrate how ML models can be used to accelerate the screening of prospective homogeneous catalyst candidates, thereby enabling the computational discovery of novel catalytic materials. To this end, we selected the problem of finding catalysts for the Suzuki–Miyaura C–C cross-coupling reaction (Fig. 1).[86–88] Specifically, we trained and applied ML models using the reaction energy associated with oxidative addition (eqn (1)), which has previously been shown to be a descriptor variable for analyzing the catalytic cycle thermodynamics using volcano plots.[39] Although kinetic profiles are obviously important for obtaining a full and accurate description of catalytic performance, here we rely on a simplified thermodynamic picture (Fig. 1), which can be exploited to rapidly discriminate between catalysts with promising or inadequate energy profiles.[39–42]

$$L_1ML_2 + (C{=}C)Br \rightarrow L_1M(Br)\,(C{=}C)L_2,\ \Delta E_{\text{Rxn A}} \qquad (1)$$

Using machine learning models of this quantity, along with previously constructed molecular volcano plots, it is possible to

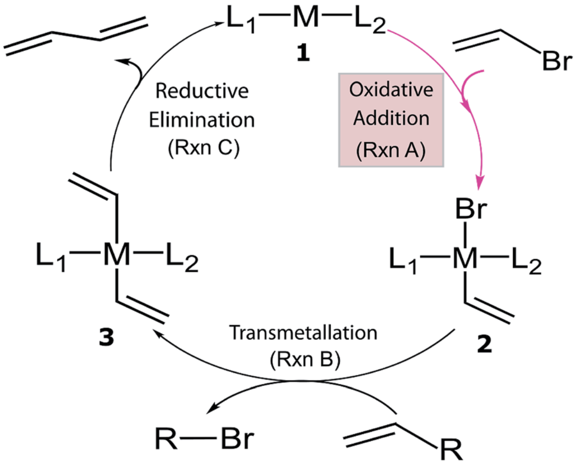

Fig. 1 General catalytic cycle for C–C cross-coupling reactions. Coupling partners (R) depend on specific cross-coupling reactions. Suzuki coupling undergoes a ligand exchange step replacing Br by an alkoxy group before transmetallation (Rxn B). The dissociated compound in Rxn B is alkoxy–R instead of Br–R and R is $[B(OH)_2(O^tBu)]^-$ for the Suzuki reaction.[39,40]

screen thousands of potential catalysts with controlled accuracy (by virtue of learning curves) and at a negligible computational overhead.

2 Computational details

The initial set of Cartesian coordinates for each catalyst was obtained by converting Simplified Molecular Input Line Entry System (SMILES) formats (*i.e.*, a line notation for entering and representing molecules and reactions)[89,90] into three-dimensional structures with the 3D structure generator operation (*i.e.*, gen3d operation) of the OpenBabel software (see the ESI† for details).[91] To generate target energy values for the training and test complexes, we proceeded as follows: computations involving geometry optimization and electronic energies were generated and executed *via* the AiiDA automated platform.[92] Gas phase geometry optimizations were computed at the B3LYP[93–95]-D3 (ref. 96 and 97) with 3-21G (for Ni, Pd, Cu and Ag complexes)[98–101] and a def2-SVP[102] basis set (for Pt and Au complexes) in Gaussian09.[103] Single point energies were computed at the B3LYP-D3/def2-TZVP level.[104] The oxidation states of the catalysts were adjusted to comply with the dominant $14e^-/16e^-$ nature of the complexes in the Suzuki cross-coupling reaction. The reaction electronic energies (eqn (1)) were computed and used as a descriptor (see a volcano plot in Fig. 2) for training the machine learning models. The ML models were trained and applied using the Quantum Machine Learning toolkit QMLcode.[105]

The reference volcano plot associated with the catalytic cycle of Fig. 1 was constructed according to the procedure outlined in our previous work[39,43] (detailed description of the procedure can be found in the ESI†) using the same theory level as for the descriptors of the machine learning training set. Note that despite the relatively modest level of theory used herein (engendered by the large computational effort associated with generating the training set for the ML model), the geometries and key energetic properties are in line with those previously computed (see Table S1†).[39,40] Similarly, we previously demonstrated that the same set of linear free-energy scaling relationships capably describe variations in the number of coordinated ligands (*i.e.*, bis *vs.* monoligated), as well as different oxidation or spin states of the catalyst.[39,42,43] Rather than predicting the entire volcano plot, the most essential property is the descriptor [ΔE(Rxn A)] (eqn (1)), which can be machine learned, as well as knowledge about its target value, *i.e.*, the energy range corresponding to the ideal plateau region (extending from −32.1 to −23.0 kcal mol^{-1}, Fig. 2).

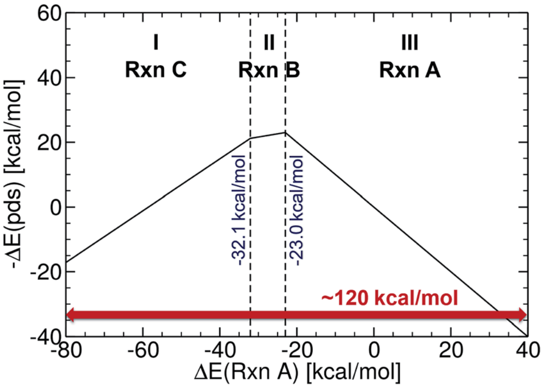

Fig. 2 Reference volcano plot for the Suzuki cross-coupling reaction. Region (I) corresponds to reductive elimination, (II) to transmetallation, and (III) to oxidative addition. Acceptable catalysts should fall into the mid region (in between −32.1 and −23.0 kcal mol^{-1}).

3 Methods

3.1 Database

The training procedure relies upon constructing a large database of catalysts that are obtained through combining various ligands and metals. These species are then used for training and testing the ML models which, in turn, are used to predict descriptor values so rapidly and with such accuracy that large libraries can be scanned in order to identify acceptable catalyst candidates. Ninety-one ligands including CO, phosphines, N-heterocyclic carbenes and pyridines were combined with six transition metals (Ni, Pd, Pt, Cu, Ag, and Au) to form the database. All possible metal/ligand combinations (*i.e.*, L_1 and L_2, where L_1ML_2 is equivalent to L_2ML_1) of catalytic cycle intermediates **1** and **2** (Fig. 1) lead to a total library consisting of 25 116 species for each intermediate (see the ESI† for a complete list of ligands used). Rather than providing the optimized structures for each ligand to build the catalysts, the geometries of catalytic cycle intermediates **1** and **2** for each database entry were created by converting SMILES strings (Fig. 3)[89,90,106] to Cartesian coordinates using the OpenBabel implementation[91] of the Merck Molecular Force Field method (MMFF94).[107–111] This database was divided into two subsets: (i) the training/test set used within cross-validated learning curves (see details on the cross-validation procedure in Section 3.2) for which the computed descriptor values [ΔE(Rxn A)] were used as a reference and (ii) the prediction set on which the model was applied to screen candidates based on their ML predicted descriptor values. Since collecting reference data for the training and test sets involves costly DFT geometry relaxations, we proceeded in two steps:[112] first, an initial training set made of complexes involving a diverse set of ligands (72 in total) with Pd (2595 complexes).[113] Secondly, a small subset of illustrative ligands (12) with each of five other metals (Pt, Au, Ag, Cu, Ni) (390 complexes) was created. The final set consisted of a total of 7054 reaction energy values corresponding to our descriptor. All DFT optimized geometries and computed electronic energies of each intermediate **1** and **2** as well as the associated ΔE(Rxn A) values are provided in the ESI.† ML models were trained on this set (*vide infra*), and out-of-sample predictions were then made on the prediction set that consisted of all the other complexes (18 062 in total). Note that included in this set are 19 realistic ligands that have already been employed in experimental settings (*i.e.*, ligand no. 72–90 in Fig. 3).[114–116]

3.2 Training

To begin the machine learning process, information intrinsically contained within each three-dimensional structure must

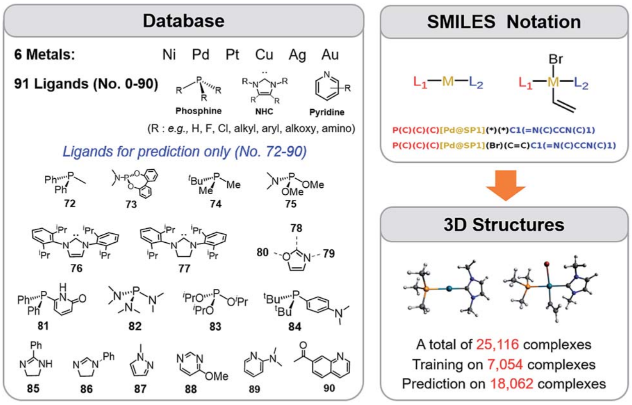
Database
6 Metals: Ni Pd Pt Cu Ag Au
91 Ligands (No. 0-90)
Phosphine
NHC
Pyridine
(R : e.g., H, F, Cl, alkyl, aryl, alkoxy, amino)
Ligands for prediction only (No. 72-90)
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
SMILES Notation
L1—M—L2
Br
L1—M—L2
P(C)(C)(C)[Pd@SP1](*)(*)C1(=N(C)CCN(C)1)
P(C)(C)(C)[Pd@SP1](Br)(C=C)C1(=N(C)CCN(C)1)
3D Structures
A total of 25,116 complexes
Training on 7,054 complexes
Prediction on 18,062 complexes

Fig. 3 A database of 25 116 molecular transition-metal catalyst candidates. Each complex consists of one out of six transition metals and a combination of two out of 91 ligands (left) (see the ESI† for details). Each ligand was written as the SMILES notation and all possible L_1–M–L_2 combinations were constructed (top right-hand corner). SMILES strings were then converted into Cartesian coordinates through the 3D structure generator of the OpenBabel software (bottom right-hand corner). DFT reference results for training and testing of ML models were obtained for a sub-set of 7054 candidates. Those structures were exploited for computing binding energies and for training the ML models.

be transformed into a suitable representation. The approach selected to represent a molecule has a crucial impact on the learning curve (for a recent example of a study discussing the role of the representation, see ref. 83). It is of particular importance to construct a meaningful relationship between the representation and the catalyst candidate, that will be learned by the machine learning algorithm. For this reason, all the relevant variables for computing the target properties (in our case ΔE(Rxn A)) should be represented in the chosen machine learning representation of the molecule. Over the last few years, increasingly improved representations[44,78,82,117,118] that progressively encode increasing amounts of physical information have been proposed. Here, we focus on the sorted Coulomb Matrix (CM), the first representation introduced for ML models trained throughout chemical space and used to predict quantum properties,[44] a two-body bagged variant of the CM with superior performance, the Bag of Bonds (BoB),[78] and the recently proposed Spectrum of London and Axilrod–Teller–Muto potential (SLATM).[119] The CM representation consists of a square atom by atom matrix, where the diagonal elements model the potential energies of free atoms while the off-diagonal elements correspond to the Coulomb nuclear repulsion between atom pairs. In the BoB representation, CM elements are bagged (*e.g.*, C–C, C–N, C–H, *etc.* are accounted for in separate bags.). SLATM is based on the dissociative limits of intermolecular dispersion contributions between unpolarized moieties. They account for interatomic two-body terms through London's dispersion curve (rather than Coulomb), and for the three-body terms according to Axilrod–Teller–Muto.[120,121]

We stress that our principal objective is to describe the oxidative addition step directly from rough-coordinate estimates obtained from the SMILES structure (*i.e.*, without providing accurate geometry as an input). After conversion from SMILES to coordinates, we map our input representation onto the corresponding continuous label value (here ΔE(Rxn A)) using kernel ridge regression (KRR),[122] which solves nonlinear problems by mapping data from the input space into a high-dimensional linear feature space (kernel trick). A Laplacian kernel function is used for the CM and BoB representations, and a Gaussian kernel for the SLATM representation (more details in the ESI†). The quality of the models is evaluated by reporting test errors, which can be obtained by separating the dataset into training and test frames and calculating the average error (typically a mean absolute error (MAE)) for the predictions on the out-of-sample test set. This random subsampling cross-validation procedure[46] was used to shuffle the dataset randomly into different training sets. For every shuffling step the MAE for the model was calculated and the procedure repeated ten times for every training set size N. Afterwards, the errors for the different models were averaged into a single cross-validated error. Note that this error remains a random variable that is dependent on the initial splitting of the training/test datasets. When plotted on a log–log scale, successful learning is indicated by linearly decaying behavior for large training set sizes, as already suggested by Vapnik and others in the nineties.[123,124]

4 Results and discussion

4.1 Machine learning

In order to verify the performance and validity of our ansatz, we have trained and tested machine learning models for various training set sizes. The resulting learning curves, depicted in Fig. 4, demonstrate the efficiency and accuracy of the learning process in terms of a near-linear decay of test error with training set size. While learning is observed for all representations, the learning curves illustrate the impact of the molecular representation on the off-set and slope. Overall, the performances of the ML models based on the SLATM and BoB are very similar (for the largest training set, the MAE is 2.61 kcal mol^{-1} and 2.73 kcal mol^{-1} respectively) and superior to CM (largest training set MAE = 3.05 kcal mol^{-1}). Despite these small variations, it is obvious that efficient learning is achieved by all three representations. This result contrasts with findings in ref. 51 where the CM was claimed to be of little use when constructing ML models for transition metal complexes. The poor performance of CM is more likely due to inappropriate choice of properties (electronic spin-states) than to the molecular systems themselves. It seems intuitive that any purely structure and composition based representation will struggle to account for various electronic states. When it comes to simple electronic ground state properties, such as the oxidative addition step studied here, Fig. 4 clearly demonstrates that the CM is very applicable to the machine learning modeling of properties of transition metal complexes. We also note that the BoB representation performs surprisingly well for this problem. We ascribe this behavior to the bagging which allows the model to place appropriate weights to bonds involving the transition metal.

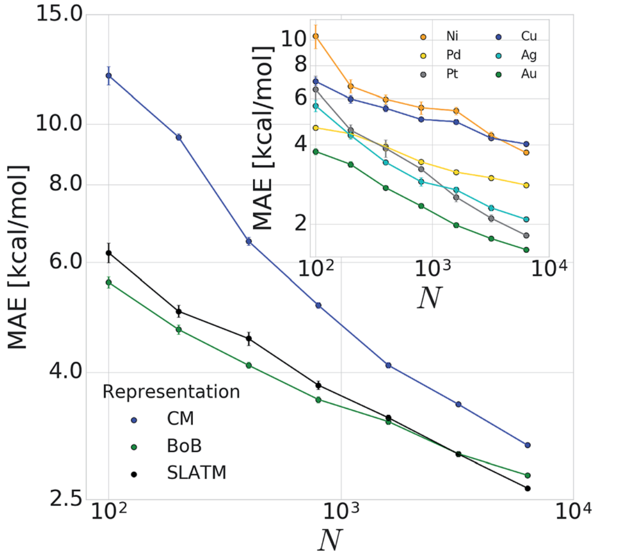

Fig. 4 Learning curves (test error of catalytic descriptor values as a function of training set size (*N*)) for oxidative addition of vinyl bromide. Error bars correspond to standard deviation in cross validation. Inset: corresponding learning curves for individual metals for BoB.

The energy range for the descriptors of the training set (corresponding to the *x*-axis of the molecular volcano plot) is ≈120 kcal mol^{-1} (Fig. 2). We therefore considered the ML model to be sufficiently well converged for the task of picking catalysts, once the learning curve dropped to less than 3 kcal mol^{-1} (*i.e.*, 2% of the descriptor range). The most efficient representations, SLATM and BoB, reached this threshold with a training set of 7054 binding energies. The following discussion will thus be based on the less sophisticated representation, BoB. All the predictions associated with the other two representations are presented in the ESI.† It is important to reiterate that while the machine learning models were trained on DFT reaction energies obtained for DFT optimized geometries, the molecular representations in the test set were constructed solely from the coordinates directly obtained from SMILES conversion.

The heterogeneity of the training set[112] (*i.e.*, unequal representation of the six transition metals) has been looked into by evaluating the individual predictions of the BoB based machine learning model on each metal separately. The resulting learning curves depicted in the inset of Fig. 4 demonstrate that learning is attained for all metals. For the largest training set size, the target MAE of 3 kcal mol^{-1} is achieved for Pd, Pt, Ag and Au, while the Ni and Cu metal complexes are less accurately described (best MAE = 3.74 and 4.04 kcal mol^{-1}, respectively). These larger errors certainly originate from the smaller sample of Ni complexes and from copper-ligand combinations featuring ligands that are less frequent in the rest of the training set. This leads to a larger energy range in the descriptor variables which can be seen as a broader distribution/width (see the histograms (Fig. S2 and S3) in the ESI†). Overall, however, the ML performance for Ni and Cu-based complexes is still useful as it is not more than 5% of the descriptor's energy range (*i.e.*, inferior to 5 kcal mol^{-1}).

4.2 Catalyst prediction

The trained ML models were subsequently exploited to predict the energy based descriptor of 18 062 potential out-of-sample catalysts with negligible computational cost (*vide supra*). At this point, it is worth noting that out-of-sample predictions that involve ligands not previously seen by the models should be considered with more care. Additionally, the predictive power of the model would be limited for catalysts that would suffer from a convergence problem in an actual computation.[113] Because we are interested only in the catalysts predicted to have the best thermodynamic profile for the Suzuki–Miyaura reaction, emphasis was placed on a narrow range of descriptor energy values (from −32.1 to −23.0 kcal mol^{-1}) corresponding roughly to the plateau of the volcano. However, the same ML models would be relevant to the analysis of other cross-coupling reaction variants differing only by the width of the plateau region.[40] Using the BoB model, 557 catalysts were identified that fell into this region. A brief examination of the metal distribution (Fig. 5) yields expected results, namely that catalysts incorporating group 10 metals (Ni, Pd, Pt) appear more frequently than their

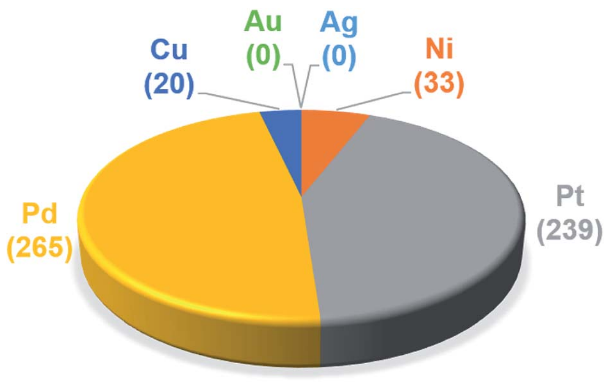

Fig. 5 Occurrence of the six metal complexes in the selected range of −32.1/−23.0 kcal mol^{-1} predicted by the machine learning model using the BoB representation.

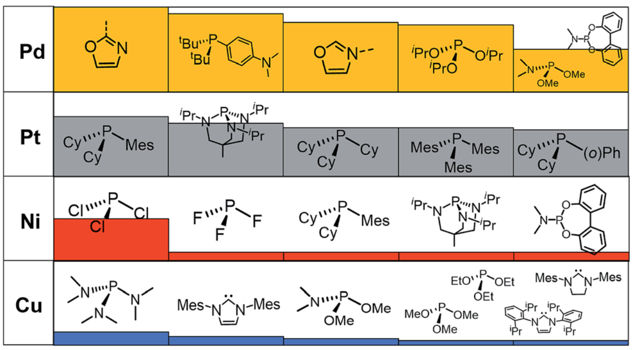

Fig. 6 Histogram ranking of the five most identified ligands that appear on the volcano plateau (*i.e.*, with descriptor values between −32.1 and −23.0 kcal mol^{-1}) by metal type as predicted by the machine learning model using the BoB representation. The histogram is scaled relative to the Pd/oxazole ligand combination, which has the highest metal/ligand occurrence appearing 38 times on the volcano plateau.

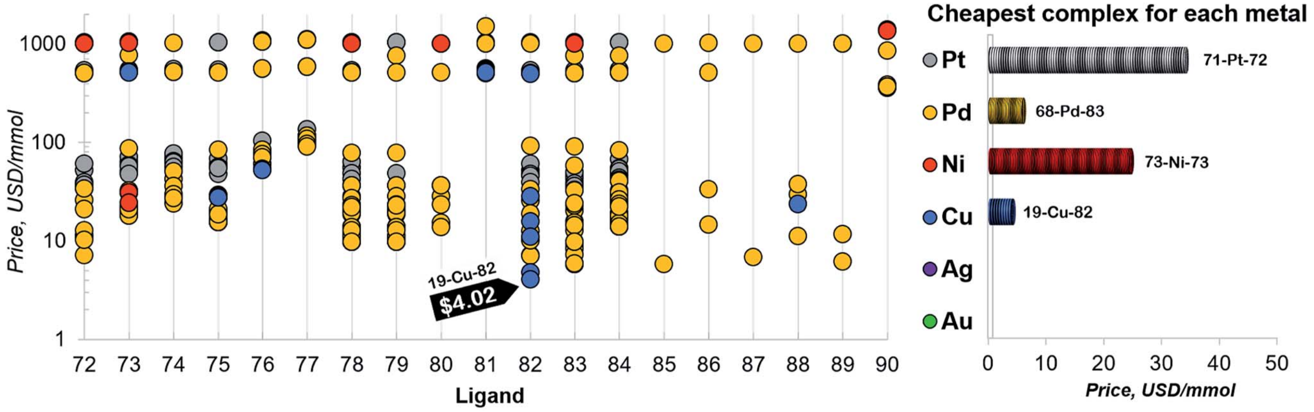

Fig. 7 Estimated price (for one mmol in US dollars) of the catalysts in the selected range of $-32.1/-23.0$ kcal mol^{-1} (for ligand no. 72–90). The price is calculated as a summation of the commercial price of transition metal precursors (one mmol) and one mmol of each ligand. The cheapest complex for each metal is shown on the right. The estimated price of all the 557 catalysts is detailed in the ESI.†

group 11 (Cu, Ag, Au) counterparts. This finding is in line with our earlier DFT-based molecular volcano plot analysis of the same reaction.[39–41]

A prevalent metal identified by the ML model is palladium, which has 265 species that appear on or near the volcano plateau (Fig. 5). The large number of Pd catalysts attests to the accuracy of the ML models, as these species have a rich history in catalyzing cross-coupling reactions.[125–128] On the other hand, Pt catalysts are virtually experimentally unknown[129] and those that have been tested tend to show only moderate catalytic ability.[130] Nonetheless, their significant presence on the volcano plateau does align with our earlier DFT-based evaluations.[39–41] Indeed, we previously postulated that the presence of Pt based catalysts on top of the volcano may indicate that the problem with these species is less thermodynamic and more kinetic in nature.[40] In addition, others have speculated that an enhanced M–R bond strength causes transmetallation in these species to be sluggish.[131] Despite being well-known cross-coupling catalysts,[132] only a handful of Ni based species are predicted by the ML model to appear near the volcano plateau. However, in its current state, the ML models consider only a single oxidation state, that for Ni corresponds to a Ni(0)/Ni(II) based catalytic cycle. Thus, the more catalytically active Ni(I) oxidation state, which is accessed *via* a one-electron redox process[133] and generally shifts Ni catalysts from the strong-binding side of the volcano onto the plateau,[42] is currently not assessed by the ML models (*vide supra*) but incorporation of alternative catalytic oxidation states represents an appealing future improvement of the current model. The volcano plot also reveals the influence of ligand type on the thermodynamics of the catalytic cycle. For example, Fig. 6 clearly indicates that phosphine ligands generally outperform N-heterocyclic carbene and pyridine ligands when combined with group 10 metal (Ni, Pd, and Pt) complexes. More interesting is the presence of oxazole ligands for Pd metals. While the use of the monodentate variant (*e.g.*, ligands no. 78–80) appears elusive in the literature, the chemistry associated with the use of bidentate bis(oxazole) ligands for cross-coupling reactions is relatively well established.[134]

By far, the vast majority of the coinage metal (group 11) catalysts have very weak binding energies and, correspondingly, lie on the right (weak-binding) slope of the volcano. Indeed, no Au or Ag based catalyst has sufficiently strong binding energy to appear on the volcano plateau (Fig. 5). This finding directly agrees with experimental and computational studies that have found Ag and Au catalysts to have unfavorable free energies associated with oxidative addition.[135] On the other hand, a handful (20) of Cu based catalysts are found to have nearly ideal thermodynamic profiles. While instances of Cu-based Suzuki coupling have appeared in the literature,[136,137] these catalysts tend to employ bidentate acetylacetone (acac) or acetate/triflate ligands.[138,139] Thus, it is interesting to note that each of the thermodynamically most appealing Cu catalysts involves either a tris(dimethylamino)phosphine or bulky N-heterocyclic carbene (Fig. 6). These findings represent a potentially interesting research direction that should be explored in more depth and that has been revealed solely through the application of ML models coupled with molecular volcano plots.

Finally, a more refined selection of catalysts was obtained based on their estimated price per mmol (Fig. 7). Among the 557 catalysts with promising thermodynamic profiles, 37 complexes have an estimated price less than 10 US$ per mmol. These species include earth abundant metals (copper with tris(dimethylamino)phosphine) and a multitude of more standard palladium phosphine combinations.

5 Conclusions

We have trained and used machine learning models to dramatically accelerate the descriptor screening of 18 062 homogeneous catalysts for the Suzuki–Miyaura C–C cross-coupling reaction. The model was based on the capability of molecular volcano plots to identify thermodynamically attractive candidates with respect to a simple energy descriptor. Overall, we have identified 37 promising low-cost complexes featuring palladium and copper combined with both standard and less expected ligands. Our findings also indicate that machine learning can be used to screen thousands of catalysts, and that previously introduced machine learning representations can be used for property predictions of transition-metal complexes. Exploitation of a Δ-machine learning approach represents an appealing future improvement of the proposed ML models.[47,117]

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The National Centre of Competence in Research (NCCR) Materials Revolution: Computational Design and Discovery of Novel Materials (MARVEL) of the Swiss National Science Foundation (SNSF) and the EPFL are acknowledged for financial support. C. C. thanks Dr Matthew D. Wodrich and Prof. Konrad Patkowski for extensive discussion and critical reading of the manuscript. Prof. Jerome Waser is acknowledged for helpful discussions.

Notes and references

1 K. D. Collins, T. Gensch and F. Glorius, *Nat. Chem.*, 2014, **6**, 859–871.

2 C. Jakel and R. Paciello, *Chem. Rev.*, 2006, **106**, 2912–2942.

3 M. T. Reetz, *Angew. Chem., Int. Ed.*, 2001, **40**, 284–310.

4 S. Senkan, *Angew. Chem., Int. Ed.*, 2001, **40**, 312–329.

5 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, *Acc. Chem. Res.*, 2016, **49**, 1292–1301.

6 A. B. Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak and S. D. Dreher, *Science*, 2015, **347**, 49–53.

7 M. R. Friedfeld, M. Shevlin, J. M. Hoyt, S. W. Krska, M. T. Tudge and P. J. Chirik, *Science*, 2013, **342**, 1076–1080.

8 D. W. Robbins and J. F. Hartwig, *Science*, 2011, **333**, 1423–1427.

9 M. S. Sigman and E. N. Jacobsen, *J. Am. Chem. Soc.*, 1998, **120**, 4901–4902.

10 M. T. Reetz, *Angew. Chem., Int. Ed.*, 2002, **41**, 1335–1338.

11 Z.-M. Chen, M. J. Hilton and M. S. Sigman, *J. Am. Chem. Soc.*, 2016, **138**, 11461–11464.

12 Z. L. Niemeyer, A. Milo, D. P. Hickey and M. S. Sigman, *Nat. Chem.*, 2016, **8**, 610–617.

13 J.-Y. Guo, Y. Minko, C. B. Santiago and M. S. Sigman, *ACS Catal.*, 2017, **7**, 4144–4151.

14 K. C. Harper and M. S. Sigman, *Science*, 2011, **333**, 1875–1878.

15 T. Sperger, I. A. Sanhueza, I. Kalvet and F. Schoenebeck, *Chem. Rev.*, 2015, **115**, 9532–9586.

16 M. G. Evans and M. Polanyi, *Trans. Faraday Soc.*, 1938, **34**, 11–24.

17 R. P. Bell, *Proc. R. Soc. London, Ser. A*, 1936, **154**, 414–429.

18 L. P. Hammett, *J. Am. Chem. Soc.*, 1937, **59**, 96–103.

19 L. P. Hammett, *Chem. Rev.*, 1935, **17**, 125–136.

20 L. P. Hammett, *Trans. Faraday Soc.*, 1938, **34**, 156–165.

21 C. B. Santiago, A. Milo and M. S. Sigman, *J. Am. Chem. Soc.*, 2016, **138**, 13424–13430.

22 J. N. Brønsted and K. J. Pedersen, *Z. Phys. Chem.*, 1924, **108**, 185–235.

23 R. Parsons, *Trans. Faraday Soc.*, 1958, **54**, 1053–1063.

24 H. Gerischer, *Bull. Soc. Chim. Belg.*, 1958, **67**, 506–527.

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

26 I. C. Man, H.-Y. Su, F. Calle-Vallejo, H. A. Hansen, J. I. Martinez, N. G. Inoglu, J. Kitchin, T. F. Jaramillo, J. K. Nørskov and J. Rossmeisl, *ChemCatChem*, 2011, **3**, 1159–1165.

27 H. Dau, C. Limberg, T. Reier, M. Risch, S. Roggan and P. Strasser, *ChemCatChem*, 2010, **2**, 724–761.

28 V. Vorotnikov and D. G. Vlachos, *J. Phys. Chem. C*, 2015, **119**, 10417–10426.

29 I. Z. Kiss, Z. Kazsu and V. Gaspar, *Phys. Chem. Chem. Phys.*, 2009, **11**, 7669–7677.

30 J. O. Bockris and T. Otagawa, *J. Electrochem. Soc.*, 1984, **131**, 290–302.

31 S. Trasatti, *Electrochim. Acta*, 1984, **29**, 1503–1512.

32 J. Greeley and N. M. Markovic, *Energy Environ. Sci.*, 2012, **5**, 9246–9256.

33 J. K. Nørskov, T. Bligaard, A. Logadottir, J. R. Kitchin, J. G. Chen, S. Pandelov and U. Stimming, *J. Electrochem. Soc.*, 2005, **152**, J23–J26.

34 Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov and T. F. Jaramillo, *Science*, 2017, **355**, eaad4998.

35 P. Sabatier, *La catalysise en chimie organique*, Librairie polytechnique, 1913.

36 S. Kozuch and S. Shaik, *Acc. Chem. Res.*, 2011, **44**, 101–110.

37 V. P. Ananikov, *Understanding Organometallic Reaction Mechanisms and Catalysis: Computational and Experimental Tools*, Wiley, 2014.

38 G. Swiegers, *Mechanical Catalysis: Methods of Enzymatic, Homogeneous, and Heterogeneous Catalysis*, Wiley, 2008.

39 M. Busch, M. D. Wodrich and C. Corminboeuf, *Chem. Sci.*, 2015, **6**, 6754–6761.

40 M. Busch, M. D. Wodrich and C. Corminboeuf, *ACS Catal.*, 2017, **7**, 5643–5653.

41 M. Busch, M. D. Wodrich and C. Corminboeuf, *ChemCatChem*, 2018, **10**, 1592–1597.

42 M. D. Wodrich, B. Sawatlon, M. Busch and C. Corminboeuf, *ChemCatChem*, 2018, **10**, 1586–1591.

43 M. D. Wodrich, M. Busch and C. Corminboeuf, *Chem. Sci.*, 2016, **7**, 5723–5735.

44 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

45 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *New J. Phys.*, 2013, **15**, 095003.

46 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.

47 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.

48 T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chem. Rev.*, 2012, **112**, 2889–2919.

49 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.

50 O. A. von Lilienfeld, *Angew. Chem., Int. Ed.*, 2018, **57**, 4164–4169.

51 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.

52 J. P. Janet, L. Chan and H. J. Kulik, *J. Phys. Chem. Lett*, 2018, **9**, 1064–1071.

53 J. P. Janet and H. J. Kulik, *Chem. Sci.*, 2017, **8**, 5137–5152.

54 A. G. Maldonado and G. Rothenberg, *Chem. Soc. Rev.*, 2010, **39**, 1891–1902.

55 E.-J. Ras, M. J. Louwerse and G. Rothenberg, *Catal. Sci. Technol.*, 2012, **2**, 2456–2464.

56 E.-J. Ras and G. Rothenberg, *RSC Adv.*, 2014, **4**, 5963–5974.

57 N. Madaan, N. R. Shiju and G. Rothenberg, *Catal. Sci. Technol.*, 2016, **6**, 125–133.

58 E. Vignola, S. N. Steinmann, B. D. Vandegehuchte, D. Curulla, M. Stamatakis and P. Sautet, *J. Chem. Phys.*, 2017, **147**, 054106.

59 J. R. Kitchin, *Nature Catalysis*, 2018, **1**, 230–232.

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

61 J. Noh, S. Back, J. Kim and Y. Jung, *Chem. Sci.*, 2018, **9**, 5152–5259.

62 J. H. Friedman, *Comput. Stat. Data Anal.*, 1999, **38**, 367–378.

63 I. Takigawa, K.-i. Shimizu, K. Tsuda and S. Takakusagi, *RSC Adv.*, 2016, **6**, 52587–52595.

64 R. Gasper, H. Shi and A. Ramasubramaniam, *J. Phys. Chem. C*, 2017, **121**, 5612–5619.

65 R. Jinnouchi and R. Asahi, *J. Phys. Chem. Lett.*, 2017, **8**, 4279–4283.

66 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.

67 M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *J. Phys. Chem. Lett.*, 2014, **5**, 3056–3060.

68 M. Fernandez, N. R. Trefiak and T. K. Woo, *J. Phys. Chem. C*, 2013, **117**, 14095–14105.

69 H. Xin, A. Holewinski and S. Linic, *ACS Catal.*, 2012, **2**, 12–16.

70 X. Ma, Z. Li, L. E. K. Achenie and H. Xin, *J. Phys. Chem. Lett.*, 2015, **6**, 3528–3533.

71 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.

72 C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, Mass, 2006.

73 Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan and J. K. Nørskov, *ACS Catal.*, 2017, **7**, 6600–6608.

74 Z. W. Ulissi, A. R. Singh, C. Tsai and J. K. Nørskov, *J. Phys. Chem. Lett.*, 2016, **7**, 3931–3935.

75 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.

76 R. B. Wexler, J. M. P. Martirez and A. M. Rappe, *J. Am. Chem. Soc.*, 2018, **140**, 4678–4683.

77 G. A. Landrum, J. E. Penzotti and S. Putta, *Meas. Sci. Technol.*, 2005, **16**, 270–277.

78 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

79 M. Rupp, R. Ramakrishnan and O. A. von Lilienfeld, *J. Phys. Chem. Lett.*, 2015, **6**, 3309–3313.

80 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.

81 F. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Int. J. Quantum Chem.*, 2015, **115**, 1094–1101.

82 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.

83 B. Huang and O. A. von Lilienfeld, *J. Chem. Phys.*, 2016, **145**, 161102.

84 T. Bereau, D. Andrienko and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 3225–3233.

85 N. J. Browning, R. Ramakrishnan, O. A. von Lilienfeld and U. Roethlisberger, *J. Phys. Chem. Lett.*, 2017, **8**, 1351–1359.

86 N. Miyaura, K. Yamada and A. Suzuki, *Tetrahedron Lett.*, 1979, **20**, 3437–3440.

87 N. Miyaura and A. Suzuki, *Chem. Rev.*, 1995, **95**, 2457–2483.

88 A. Suzuki, *Angew. Chem., Int. Ed.*, 2011, **50**, 6722–6737.

89 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Model.*, 1989, **29**, 97–101.

90 D. Weininger, *Proc. Edinb. Math. Soc.*, 1970, 1–14.

91 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.

92 G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, *Comput. Mater. Sci.*, 2016, **111**, 218–230.

93 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

94 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.

95 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

96 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

97 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.

98 R. Ditchfield, W. J. Hehre and J. A. Pople, *J. Chem. Phys.*, 1971, **54**, 724–728.

99 J. S. Binkley, J. A. Pople and W. J. Hehre, *J. Am. Chem. Soc.*, 1980, **102**, 939–947.

100 M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *J. Am. Chem. Soc.*, 1982, **104**, 2797–2803.

101 W. J. Pietro, M. M. Francl, W. J. Hehre, D. J. DeFrees, J. A. Pople and J. S. Binkley, *J. Am. Chem. Soc.*, 1982, **104**, 5039–5048.

102 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

103 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone,

G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09, Revision D.01*, Gaussian, Inc., Wallingford CT, 2016.

104 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

105 A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K. R. Müller and O. A. von Lilienfeld, *QML: A Python Toolkit for Quantum Machine Learning, v0.3.1*, 2017, DOI: 10.5281/zenodo.817332.

106 The *trans* isomerism constraint was imposed using the general chiral specification syntax of the SMILES notation (*i.e.*, the @SP square-planar class symbol) as depicted (on the top right-hand corner) in Fig. 3.

107 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

108 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 520–552.

109 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 553–586.

110 T. A. Halgren and R. B. Nachbar, *J. Comput. Chem.*, 1996, **17**, 587–615.

111 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 616–641.

112 To refine the accuracy of the ML model in the targeted descriptor energy range, *i.e.*, the top of the volcano, we exploited the trained model to predict the binding energies on a subset of complexes combining the 5 metals (Pt, Au, Ag, Cu, and Ni) and 72 ligands (from no. 0 to 71) and selected the molecules for which the ML predicted reaction energy was in the selected range (as opposed to randomly selecting additional candidates to extend the training set).

113 Due to convergence problems, exactly 2595 binding energies from Pd complexes were used in the training set.

114 P. Lei, G. Meng, Y. Ling, J. An and M. Szostak, *J. Org. Chem.*, 2017, **82**, 6638–6646.

115 R. Martin and S. L. Buchwald, *Acc. Chem. Res.*, 2008, **41**, 1461–1473.

116 S. S. David and L. B. Stephen, *Angew. Chem., Int. Ed.*, 2008, **47**, 6338–6361.

117 A. P. Bartok, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csanyi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.

118 F. A. Faber, A. S. Christensen, B. Huang and O. A. von Lilienfeld, *J. Chem. Phys.*, 2018, **148**, 241717.

119 B. Huang and O. Anatole von Lilienfeld, ArXiv e-prints, 1707.04146, 2017.

120 B. M. Axilrod and E. Teller, *J. Chem. Phys.*, 1943, **11**, 299–300.

121 Y. Muto, *J. Phys. Soc. Jpn.*, 1943, **17**, 629.

122 R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001.

123 V. Vapnik, *The nature of statistical learning theory*, Springer science & business media, 2013.

124 C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik and J. S. Denker, *Advances in Neural Information Processing Systems*, 1994, pp. 327–334.

125 A. de Meijere, S. Brase and M. Oestreich, *Metal-Catalyzed Cross-Coupling Reactions and More*, Wiley-VCH, Weinheim, 2014.

126 T. Colacot, *New Trends in Cross-Coupling: Theory and Applications*, The Royal Society of Chemistry, Cambridge, 2015.

127 Y. Nishihara, *Applied Cross-Coupling Reactions*, Springer-Verlag, Berlin, 2013.

128 G. A. Molander, *Cross-Coupling and Heck-Type Reactions*, Thieme, Stuttgart, 2013.

129 R. B. Bedford, S. L. Hazelwood and D. A. Albisson, *Organometallics*, 2002, **21**, 2599–2600.

130 C. Mateo, C. Fernandez-Rivas, D. J. Cardenas and A. M. Echavarren, *Organometallics*, 1998, **17**, 3661–3669.

131 V. P. Ananikov, D. G. Musaev and K. Morokuma, *Organometallics*, 2005, **24**, 715–723.

132 F.-S. Han, *Chem. Soc. Rev.*, 2013, **42**, 5270–5298.

133 S. Z. Tasker, E. A. Standley and T. F. Jamison, *Nature*, 2014, **509**, 299–309.

134 D. Zhang and Q. Wang, *Coord. Chem. Rev.*, 2015, **286**, 1–16.

135 M. Livendahl, C. Goehry, F. Maseras and A. M. Echavarren, *Chem. Commun.*, 2014, **50**, 1533–1536.

136 C. Maaliki, E. Thiery and J. Thibonnet, *Eur. J. Org. Chem.*, 2017, **2**, 209–228.

137 S. Thapa, B. Shrestha, S. K. Gurung and R. Giri, *Org. Biomol. Chem.*, 2015, **13**, 4816–4827.

138 H. S. P. Rao and A. V. B. Rao, *J. Org. Chem.*, 2015, **80**, 1506–1516.

139 M. Hoshi, N. Kawamura and K. Shirakawa, *Synthesis*, 2006, **12**, 1961–1970.