**Article:**

Mace, S., Xu, Y. and Nguyen, B.N. orcid.org/0000-0002-0254-025X (2024) Automated transition metal catalysts discovery and optimisation with AI and Machine Learning. ChemCatChem, 16 (10). e202301475. ISSN 1867-3880

https://doi.org/10.1002/cctc.202301475

ChemCatChem

The European Society Journal for Catalysis

Chemistry Europe

European Chemical Societies Publishing

Supported by GECATS GERMAN CATALYSIS SOCIETY

**Title:** Automated transition metal catalysts discovery and optimisation with AI and Machine Learning

**Authors:** Samuel Mace, Yingjian Xu, and Bao N. Nguyen

**To be cited as:** *ChemCatChem* **2024**, e202301475

**Link to VoR:** https://doi.org/10.1002/cctc.202301475

WILEY-VCH

10.1002/cctc.202301475

# Automated transition metal catalysts discovery and optimisation with AI and Machine Learning

Samuel Mace,[a] Yingjian Xu,*[b] Bao N. Nguyen,*[a]

*Dedicated to Prof. John M. Brown (FRS) on his 84th birthday*

Significant progress has been made in recent years in the use of AI and Machine Learning (ML) for catalyst discovery and optimisation. The effectiveness of ML and data science techniques was demonstrated in predicting and optimising enantioselectivity and regioselectivity in catalytic reactions through optimisation of the ligands, counterions and reaction conditions. Direct discovery of new catalysts/reactions is more difficult and requires efficient exploration of transition metal chemical space. A range of computational techniques for descriptor generation, ranging from molecular mechanics to DFT methods, have been successfully demonstrated, often in conjunction with ML to reduce computational cost associated with TS calculations. Complex aspects of catalytic reactions, such as solvent, temperature, etc., have also been successfully incorporated into the ML optimisation and discovery workflow.

## Introduction

Automated chemical space exploration with the help of AI/Machine Learning (ML) is a highly important methodology in modern chemical discovery. Progresses in this area with organic compounds have resulted in the first AI discovered Active Pharmaceutical Ingredient (API) entering Phase II trials.[1,2] The same benefits can be extended to catalyst discovery through chemical space exploration of organometallic compounds. However, this is significantly more challenging due to the additional constraints, *e.g.* coordination geometry, and complexity, *e.g.* spin state, catalyst stability and selectivity, etc. Evaluating the desired function of catalysts for *in silico* screening is also more computationally demanding compared to API discovery, due to the need to calculate and/or estimate properties of excited states and transition states. In homogeneous catalysis, additional dimensions such as solvent, temperature and additives can have a significant impact on reaction outcome and need to be included in the evaluation methodology. Synthetic catalytic reactions often involves chemo- and stereoselectivity, competing side reactions, and multiple possible mechanistic pathways,[3–9] depending on the substrate and catalyst (Figure 1).[10–13]

These complex and demanding challenges led to the need for of AI/ML models which can predict catalytic activity. This approach can be particularly powerful for complex and difficult substrates, which tend to occur in high value chemical synthesis. Unfortunately, experimental data on catalytic activity in this area is scarce, with the majority of the literature containing reaction yields instead of reaction rates (due to the cost of labour and resources for collecting kinetic data).[14] While successes have been reported with statistical analysis of small datasets,[15,16] the demand for training data for advanced AI/ML models necessitates accurate and low-cost molecular modelling tools for data generation.

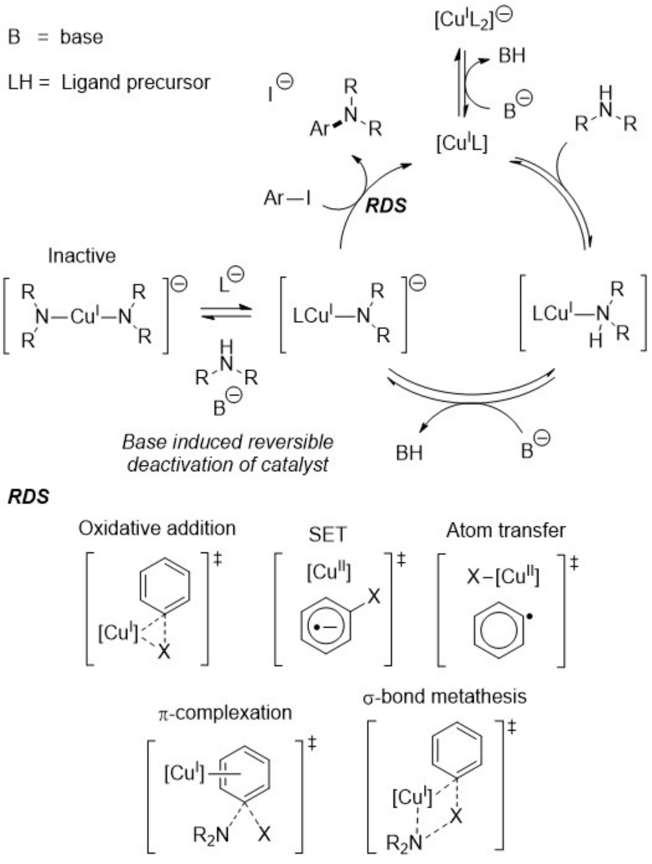


**Figure 1.** An example mechanism of the Ullmann-Goldberg coupling reaction, with a reaction condition dependent deactivation pathway and multiple possible mechanisms for the rate determining step (RDS).[17]

[a] Samuel Mace, Dr. Bao N. Nguyen*
Institute of Process Research & Development
School of Chemistry
University of Leeds
Woodhouse Lane, Leeds
LS2 9JT, United Kingdom
E-mail: b.nguyen@leeds.ac.uk

[b] Dr. Yingjian Xu*
GoldenKeys High-tech Materials Co., Ltd.
Building 3, Guizhou Industrial Investment Technology Industrial Park
Gui'an New District, Guizhou Province
550008 China Email: goldenkeys9996@thegoldenkeys.com.cn

This is an unique area of research which requires advancement in both cheminformatics and high throughput molecular modelling. A typical workflow for screening catalytic candidates will start with either experimental or computational data of a relatively small set of ligands/catalysts. This dataset is then used to train a ML model to predict the desired catalytic properties and performance (Figure 2). The model can then be used to extrapolate the performance of a much larger set of ligands/catalysts generated *in silico.*

Excellent and recent reviews from practitioners in the field have discussed predicting organic reactivity with ML,[18] general computational discovery of transition metal complexes,[19] descriptors for ML in catalysis,[20], quantum methods for computational catalysis,[21] mechanism-based models,[22] practical tutorial with code,[23] and a road-map on machine learning in electronic structure.[24] In this review, we will focus on recent peer-reviewed publications since 2020 in AI/ML-enabled organometallic catalyst discovery and optimisation, tackling the following challenges: (i) automated exploration of ligand space; (ii) computational methods for data generation; and (iii) dealing with complex aspects of catalysis such as selectivity, reaction conditions and competing pathways. Other exciting approaches based on data science, e.g. volcano plots,[25–27] or process optimisation will not be included.

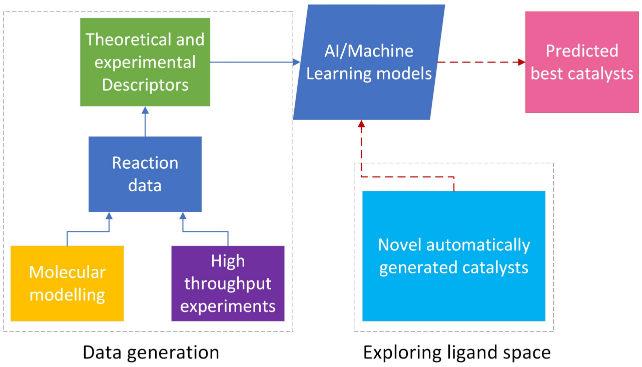

Theoretical and experimental Descriptors
AI/Machine Learning models
Predicted best catalysts
Reaction data
Molecular modelling
High throughput experiments
Novel automatically generated catalysts
Data generation
Exploring ligand space

**Figure 2.** A typical workflow for ML-guided catalyst discovery and optimisation

## Automated exploration of ligand space

Chemical space exploration is a cornerstone of AI/ML-guided chemical discovery. In the context of catalyst, it is essential for exploring both the ligand space/prediction stage (Figure 2) and the data generation stage if the data is generated computationally. For catalyst optimisation, limited exploration of similar chemical space may be sufficient. However, catalyst discovery requires highly efficient sampling of a wider chemical space or a closed-loop optimisation approach to chemical space sampling. Ultimately, this needs to be balanced with the synthetic viability of the generated catalysts and ligands to ensure their expeditious experimental validation. While the field is still far from achieving these lofty goals, recent progresses have shown these are attainable if expertise in cheminformatics can be leveraged.

The term "ligand space" has previously been used by Fey and co-workers to describe the relative positions of phosphine and C/N/O ligands in Principal Component Analysis maps based on their electronic and steric descriptors.[28–33] In the context of AI/ML-guided exploration, a more structurally oriented concept is normally adopted. This is the result of cheminformatics tools, often employed in building the structure of ligands and catalysts *in silico*. Nevertheless, the advanced techniques for exploring chemical space in medicinal chemistry, *e.g.* variations of autoencoders,[34–37] have yet to be adapted for catalysis. There are obvious difficulties associated with this, particularly maintaining and varying coordination numbers, geometry of the complexes, and the oxidation state and spin state of the metal centre. Instead, a simpler high throughput combinatorial approach has been adapted by most researchers in the field.

Kulik group developed an approach which uses a small number of ligands (typically <1000) with reasonably high symmetry to generate training data. The new ligands and complexes (>1M) are then generated from combinations of structural fragments of these training complexes. Properties were predicted using Artificial Neural Networks (ANNs) models and achieving a Mean Absolute Error (MAE) = 4.5-6 kcal·mol$^{-1}$ for $\Delta E_{HAT}$ and $\Delta E_{release}$, the barriers for the two steps in the catalytic cycles (Figure 3).[38] Macrocyclic ligands are particularly suitable for this approach as they do not have multiple suitable conformers for coordination. This was further expanded into the concept of ligand additivity by inferring heteroleptic properties from a stoichiometric combination of homoleptic complexes, which led to an interpolation scheme, which includes *cis* and *trans* isomer effects. The interpolated adiabatic high-spin to low-spin splitting (as a weighted average of the spin splitting of the parent homoleptic complexes) and HOMO energy. $\Delta E_{H-T}$, was found to match with the DFT derived values (B3LYP/LACVP*, LANL2DZ effective core potential for transition metals and the 6-31G* basis for all other atoms) for Fe(II) complexes with pairs of any of the three ligands: $CH_3CN$, $H_2O$, and CO, giving MAEs up to 2.6 kcal·mol$^{-1}$ and 0.11–0.25 eV, respectively.[39]

A different study by Gensch, Sigman and Aspuru-Guzik employed the same combinatorial approach to generate and predict properties of >300000 monophosphine ligands.[40] DFT descriptors, calculated with PBE0(D3BJ)/def2–TZVP//PBE0(D3BJ)/6–31+G(d,p) method, were generated for 1558 ligands, which were subsequently used to train highly accurate machine learning models to predict properties of new monophosphine ligands (Figure 5). This led to *kraken*, a discovery platform of 190 physicochemical descriptors for monodentate phosphine ligands (https://kraken.cs.toronto.edu). Importantly, all thermally accessible conformers of the ligands were considered, due to their non–chelating nature. *kraken* was then used to select a set of 32 commercially available ligands that samples the entire covered chemical space evenly.[41] This was achieved through dimensionality reduction of 190 condensed descriptors per ligand (78 descriptors for each conformer including Boltzmann weight average of the highest and lowest value of each property across all conformers). *k*-Means clustering algorithm, which clusters ligands with similar features together, in 4D space was used to select a diverse set of ligands for screening, leading to identification of the optimal ligands in Suzuki-Miyaura coupling reactions of aryl chlorides and aryl triflates, *i.e.* highest yields. It is worth noting that depending on how the descriptors are derived, relative position of ligands to each other in their chemical space can be significantly dif-

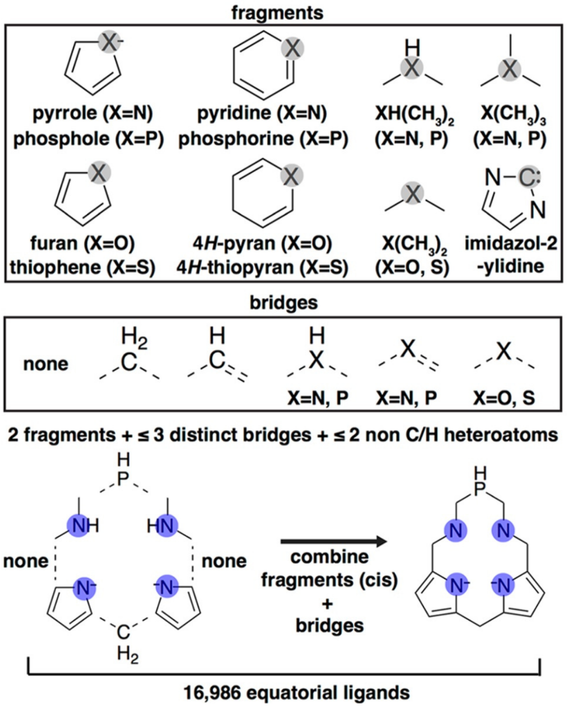


**Figure 3.** Combinatorial approach by Kulik group to explore methane-to-methanol catalyst based on porphyrin ligands (reprinted with permission from ACS)[38]

ferent.[42] This highlights the need for a consistent approach to featurisation of ligands, particularly within a single class, *e.g.* monophosphine, diphosphine, or salen ligands.

The main limitation in these early successes is the lack of quantification of transition metal chemical space. Building predictive ML models trained on relatively small datasets presents uncertainties upon extrapolation into wider chemical spaces. Recent work in Kulik group investigated how to quantify uncertainty in their ML models,[43,44] and to define distance in chemical space through a set of 25 mixed continuous and discrete features around the metal centre (*i.e.* MCDL–25).[44,45] ANN models based on these descriptors achieved a root-mean-square error (RMSE) of around 3 kcal·mol$^{-1}$ in predicting spin-state energies, $\Delta E_{H-L}$, against DFT values (B3LYP). Feature sets such as nuclear charge, electronegativity, or covalent radius on the molecular graph, which are geometry-free, have been found to be even more effective than MCDL–25 in accurately predicting redox and ionisation potential, spin–state–dependent metal-ligand bond length, and $\Delta E_{H-L}$ for transition metal complexes (Figure 4).[45] However, the performance of ANN models (used to explore a space of 5600 complexes, of which 2% was used as training data), predictably deteriorates for complexes with high feature–space distances to training data.

Coordination number and geometry are important aspects of transition metal space, particularly in catalysis. Analysis by Kulik based on >240000 mononuclear complexes in the Cambridge Structural Database (CSD) showed that approximately one third of them are octahedral and often contain monodentate ligands which can dissociate to leave empty coordination sites for catalysis.[47] Thus, the authors proceeded with a design of square-planar tetradentate ligands, leaving two coordination sites for labile monodentate ligands. This went against conventional mecha-

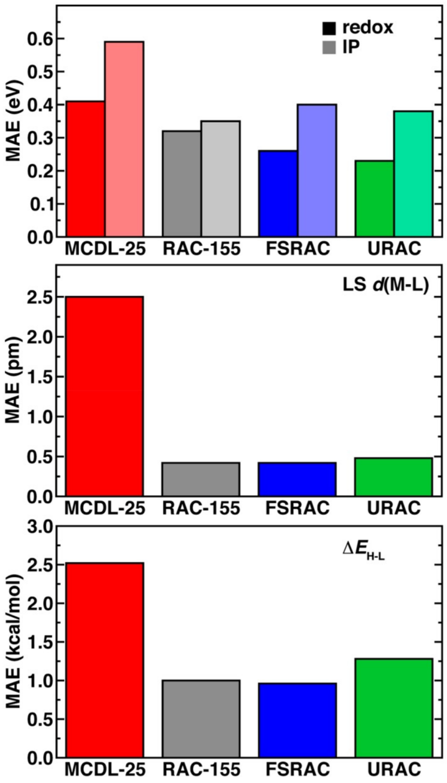


**Figure 4.** Mean absolute error (MAE) for (top) redox and ionisation potential in eV, (middle) low-spin (LS) metalligand bond length in pm, and (bottom) EHL in kcal·mol$^{-1}$. Comparisons are for the MCDL-25/ANN from ref[44] along with KRR models trained with RAC-155, a feature-selected (FS) RAC subset for each property from ref[46], and the best–overall–performing "universal" URAC 26 feature set in ref[46]. These results highlight how the systematic RAC-155 outperforms *ad hoc* MCDL-25. (Reprinted with permission from ACS)[45]

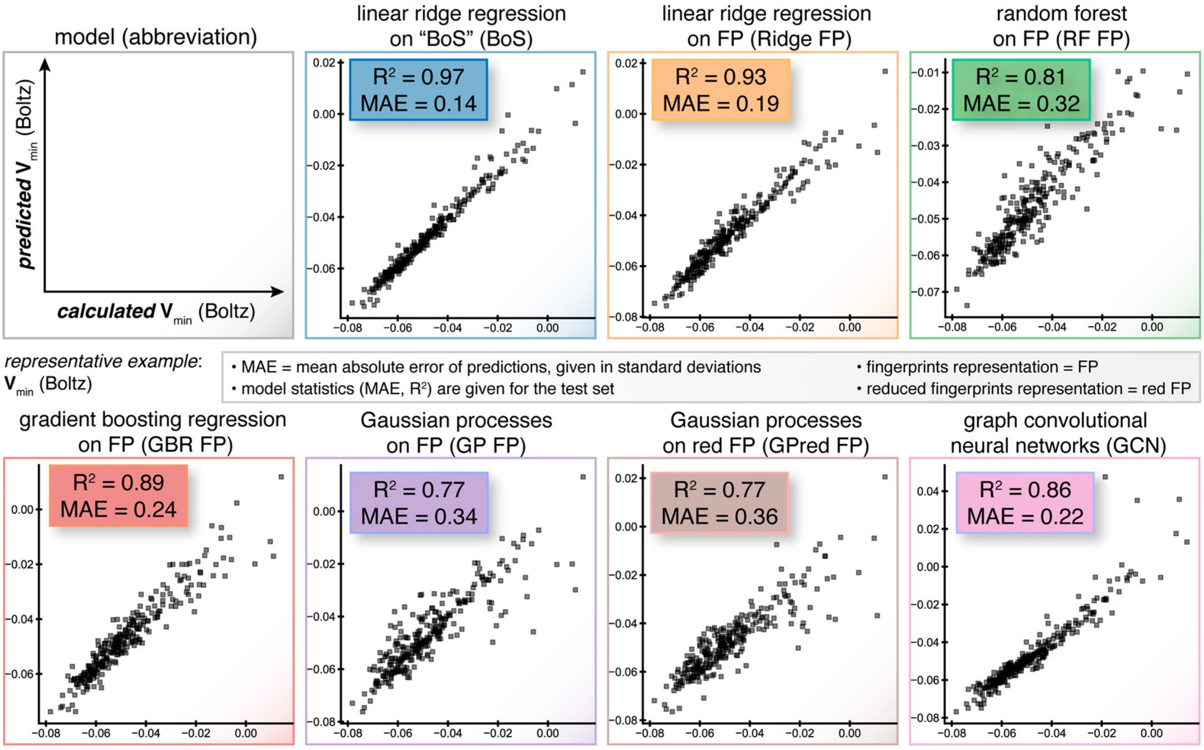


**Figure 5.** Regression performance of machine learning models. Illustrative performance of all seven types of ML models from this study for the prediction of Vmin (Boltz). BoS = Bag of Substituents; FP = fingerprint representation: circular fingerprints, radius = 2, folded to 1024 dimensions; red FP = reduced fingerprints representation: 100 most important fingerprint dimensions based on the feature importance of the GBR FP model. (Reprinted with permission from ACS)[40]

nistic understanding of metal catalysed coupling reactions. However, the targeted reactions in this case were fundamentally different, typically involving redox or electro/photochemical processes.

A possible solution for inefficient coverage of transition metal chemical space is through the use of active learning. Kulik group applied this approach to discover $3d^6$ Fe(II)/Co(III) chromophores. A consensus in predictions among 23 DFT methods across "Jacob's ladder", an established order of DFT methods with increasing accuracy and computational cost, benchmarked for large datasets of organic compounds, was used (from BP86 to DSD-PBEP6-D3BJ).[48,49] An algorithm which sampled new chemical space for additional training data based on each iteration of the prediction model led to efficient optimisation in 2D-chemical space, based on Δ–SCF gap and multi-reference character (Figure 6). Candidates with high likelihood (*i.e.* >10%) of being a chromophores were used to validate and retrain the ML models so the ML models were actively improved, leading to a 1000-fold acceleration compared to random sampling.[50] This led to the identification of Co(III) complexes with large, strong-field ligands with more saturated bonds as potential transition-metal chromophores.

Nguyen group adopted a different approach to leverage the wide chemical space covered by the Cambridge Structural Database (CSD), using both organic and organometallic structures, and to avoid the question of synthetic viability of the ligands.[51] Two closed-shell TS of the rate-determining-step (RDS) of the Ullmann-Goldberg coupling reactions were optimised with DFT (DLPNO–CCSD(T)/def2–TZVP). These TS structures were used as template to create the *catalophores* with the desired geometry and empty space

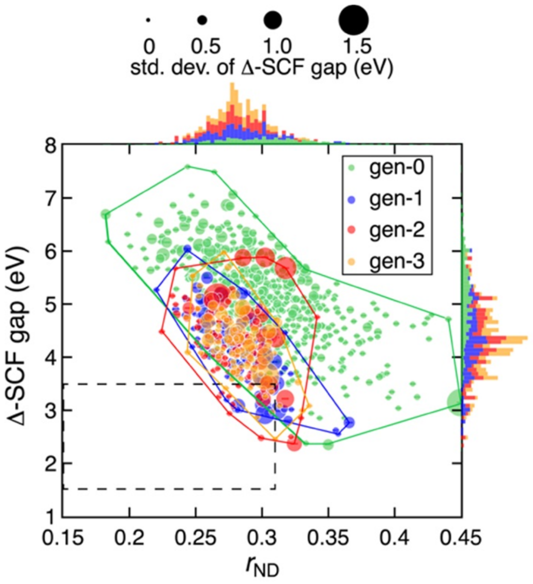


**Figure 6.** DFT-computed $r_{ND}$ vs Δ-SCF gap for base complexes in gen-0 to gen-3. For each complex, the average Δ-SCF gap over all DFAs is shown as a circle sized by the corresponding standard deviation (std. dev.) over all DFAs. The range of values sampled in each generation is indicated by a convex hull. The target zone is shown as a rectangle with dashed lines. Normalized stacked marginal histograms for Δ-SCF gap and $r_{ND}$ are also shown. (Reprinted with permission from ACS)[50]

10.1002/cctc.202301475

for the Cu(I) cation and the substrates. Searching the CSD with these *catalophores* identified 32000 of possible ligands (Figure 7). Their corresponding $\Delta G^{\ddagger}$ values (TPSS/def2-TZVP//GFN2-xT) were used to develop ML models that can predict $\Delta G^{\ddagger}$ values based on non-TS-related descriptors. The best models, using ExtraTrees and Scaled Vector Machine algorithms, gave RSME = 3.5–6.0 kcal·mol$^{-1}$ and 75-87% of the predicted $\Delta G^{\ddagger}$'s within ± 4.0 kcal·mol$^{-1}$ of the DFT values (the accuracy limit of the training data against those calculated using DLPNO-CCSD(T)/def2–TZVP).

Lastly, the efficient exploration of chemical space for organometallic catalysts requires the development of cheminformatics tools which can build and modify complexes in 3D, compared to 2D tools based on only connectivity for organic compounds. *MolSimplify* has seen widespread use for this purpose, despite its original design for geometrically rigid ground state complexes.[52,53] Modifications has been made to *MolSimplify* to enable it to build TS with asymmetrical geometry and unusual coordination numbers.[51] Alternatively, the CSD Python tool is also a highly flexible tool and for our own purposes used it to build organometallic complexes.[54] The CSD python API loads the molecule as a class object and can build and edit molecules, such as adding, removing bonds and atoms, and normalising charges and hydrogens. The API can be used to add a metal centre to ligands downloaded from CSD CrossMiner for rapid building of organometallic compounds.[55]

# Computational methods for data generation

The accuracy of any ML predictive model is ultimately limited by the quality of the training data, and while experimental data are highly valuable, currently, truly large volume of data is only accessible computationally. Thus, the quality of experimental data and the "rung" on "Jacob's ladder" of the DFT method for data generation are an integral part of developing AI/ML–assisted catalytic workflow. The larger the training dataset, the more limits are placed on the DFT method. Usually, lower level molecular modelling methods are employed to generate their training and validation data, after benchmarking against higher level DFT methods to establish their accuracy. This is even more challenging when TS properties are used as training data,[51] as optimisation of TS demands much more CPU time than optimisation of stable intermediates, ligands and starting materials. Furthermore, optimisation of TS is also more prone to errors and failures, which can lead to large amount of unproductive CPU time. On the other hand, predicting TS properties based on those of intermediates can be difficult, as famously demonstrated in the case of rhodium–catalysed asymmetric hydrogenation.[56]

In this context, the study of Balcells and Aspuru-Guzik, which used ML algorithms to predict $\Delta G^{\ddagger}$ for oxidative addition of Ir-complexes to $H_2$ directly is intriguing.[57] Instead of descriptors derived from molecular modelling, full autocorrelation features, which represented the connectivity and atomic properties (electronegativity, atomic number, coordination number and size) of the atoms at the centre of analogues of Vaska's complex, were employed. The results were benchmarked against values calculated with PBE/def2-SVP for 1947 TS, obtaining the best MAE = 1.74 kcal·mol$^{-1}$ using a Deep Neural Network (four layers with 584, 94, 41 and 20 neurons, respectively) with very significant reduction of computational time once the model is trained.

The Transition State Force Field (TSFF) technique, developed by Wiest and Norrby using the quantum-guided molecular mechanics (Q2MM) method,[58,59] is another approach which targets computational cost. It leverages very fast and computationally inexpensive force field calculations to model transition states, which traditional force fields are not capable of doing. The TSFF needs to be developed/trained for each specific reaction, based on DFT generated descriptors of a small number of reactions. The descriptors required to train the models are geometry related: bond lengths, bond angles and torsion angles. Wiest and co–workers employed TSFF models to predict enantioselectivity of different palladium catalysts in an asymmetric redox relay Heck reaction, through the stereodetermining migratory insertion step, with $R^2$ = 0.89 and Mean Unsigned Error (MUE) = 1.8 kJ·mol$^{-1}$ against 151 experimentally determined stereoselectivities (Figure 8).[60] The preferred absolute stereochemistry was correctly predicted in every case, suggesting the use of TSFF for rapid prediction of absolute stereochemistry for a class of reactions. Their TSFF model was trained on 12 separate transition states (M06–GD3/LANL2DZ(Pd) or 6-31+G*(other atoms)). Importantly, conformational search was carried out and the predicted stereoselectivity was calculated from the Boltzmann averaged conformations. Analysis of a small set of outliers linked the poor predictions to the unsatisfactory representation of $\pi$–stacking in the underlying MM3* force field.

Virtual Chemist is a software platform developed by Norrby and Moitessier, which employ Q2MM or Asymmetric Catalyst Evaluation (ACE) to predict stereoselectivity in asymmetric catalysis.[61–64] Due to its speed, four different usages were proposed: one-by-one design, library screening, hit optimisation and substrate scope evaluation. The organic catalyst candidates can be screened in hours and accuracies within 1.0 kcal·mol$^{-1}$ for $\Delta G^{\ddagger}$, although the tool has not been demonstrated with transition metal catalysts.

In spite of the low cost of TSFF, DFT methods are much more frequently employed in generating data for AI/ML applications in catalysis. Semi-empirical methods have also found extensive use in pre-optimisation of complexes and transition states. Buttar successfully used GFN2-xTB for conformational sampling of TS for 449 $S_N$Ar reactions, before optimisation with $\omega$b97xd/6-31+G(d) with SMD solvation model with MAE = 2.93 kcal·mol$^{-1}$ against experimental reaction rates.[65] In this case, a hybrid mechanism-based Gaussian Process Regression (GPR) model using reaction rate data, using B3LYP/6-31+G(d) generated descriptors, predicted $\Delta G^{\ddagger}$ values with $R^2$ = 0.87 and MAE = 0.80 kcal·mol$^{-1}$, performing better than $\omega$b97xd/6-31+G(d) in predicting $\Delta G^{\ddagger}$.

The accuracy of DFT training data, commonly accepted to be about 2–3 kcal·mol$^{-1}$,[66] is an obvious limitation of using computational data for AI/ML models. *Ab initio* methods, such as CCSD(T), give much higher accuracy, but with prohibitive computational costs for high throughput calculations. Tuckerman, Müller and Burke reported a way to overcome this,[66] using ML to calculate coupled-cluster energies from DFT densities with errors below 1.0 kcal·mol$^{-1}$ on the MD17 dataset (1500 geometries and energies of small molecules in gas phase).[67–69] This approach, known as Δ-DFT, reduces the amount of training data re-

10.1002/cctc.202301475

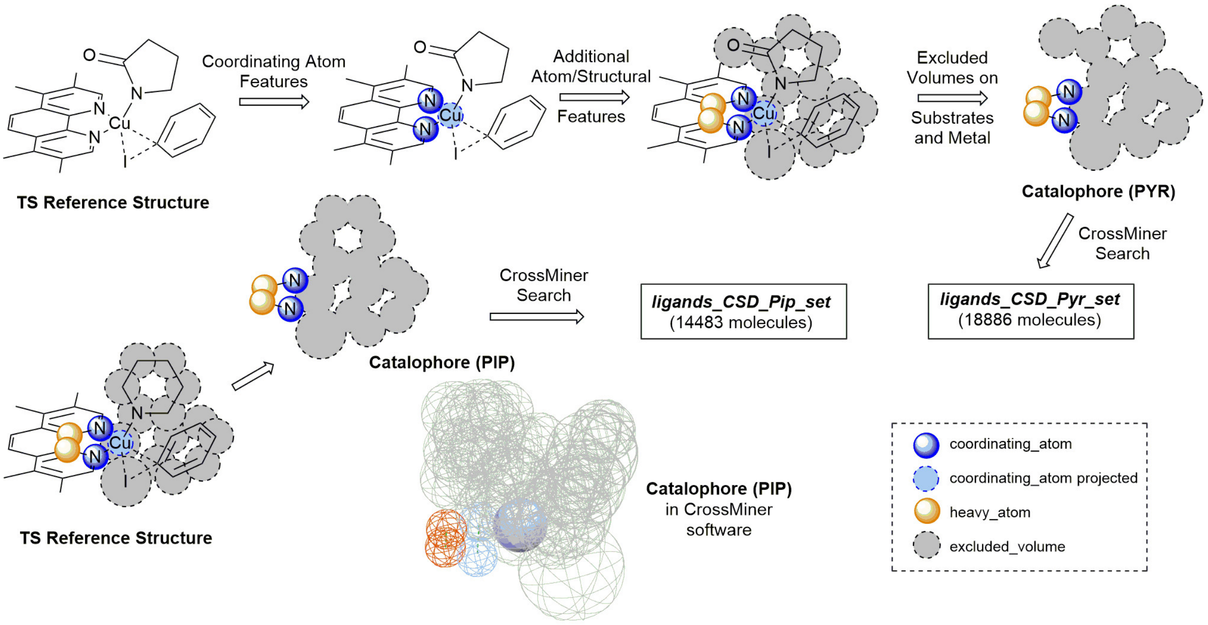


**Figure 7.** Workflow for generation of a catalophore from a transition state reference structure and identification ligands in the CSD to generate ligand sets ligands_CSD_Pip_set and ligands_CSD_Pyr_set. Hydrogens are excluded for clarity. (Reprinted with permission from RSC)[51]

quired. This allows for accurate DFT-based molecular dynamics simulations even in cases where standard DFT methods fail and may see wider application in catalysis in the future.

The combination of B97-3c//GFN2-xTB techniques, *i.e.* optimisation with GFN2-xTB followed by energy calculation with B97-3c, has been developed by Grimme for high throughput optimisation of transition metal complexes.[70,71] The method was extended to intermediates and TS in Cu(I)–catalysed Ullmann-Goldberg coupling reactions by Nguyen group.[51] Restriction on the atoms coordinating to Cu(I) was required for pre-optimisation, before successful optimisation of the TS with each ligand. The $\Delta G^{\ddagger}$ values obtained with B97-3c//GFN2-xTB were found to have a MAE = 3.9 kcal·mol$^{-1}$ against those calculated with DLPNO–CCSD(T)/def2-TZVPP (Figure 9). This level of noise in the training data compared well with predictions made by ML models using descriptors based on the catalytic intermediates before the TS: RMSE = 3.5 – 6.1 kcal·mol$^{-1}$ and 73-88% of predictions within ± 4 kcal·mol$^{-1}$ of the DFT calculated $\Delta G^{\ddagger}$. Importantly, the ML models avoid DFT optimisation and energy calculation of new TS, reducing the CPU time by up to 90%. Thus, higher level DFT methods, *e.g.* TPSS/def2-TZVP//GFN2-xTB or PBE0/def2-TZVP//GFN2-xTB, were used to generate descriptors for the ML models, leading to significant improvement to their accuracy while still reducing the CPU time by a factor of 4 when compared to direct calculation of $\Delta G^{\ddagger}$ through optimising the TS. A similar approach was employed by Ess and co-workers to automate the building and optimisation of TS for Pt-catalysed C–H activation of methane.[72] A Random Forrest (RF) model, which predicted the $\Delta G^{\ddagger}$ value, was built based on DFT data of 900 TS (PBE0-D3/Def2-SVP), but did not perform well with a validation set ($R^2 = 0.29$).

Lastly, optimisation of all the TS with DFT is prone to failure, particularly in high throughput mode. Moreover, exploration of ligand space may lead to unsuitable catalysts, for which TS optimisation may correctly fail. Thus, a significant amount of CPU time may be wasted on these jobs. Kulik group solved this problem by introducing a dynamic classifier which monitors geometry optimisation on the fly and terminates those which it predicts to be unproductive.[73,74] This classifier is based on a convolutional neural network, and makes decisions based on the evolving geometric and electronic structure and features such as energy gradient and Mulliken bond orders. This approach led to >50% reduction in CPU time while having negligible false-negative predictions (<2%) for 300 potential Mn/Fe catalysts for oxidation of methane to methanol.

## Complex aspects of catalysis

Practical protocols for catalytic reactions include temperature, solvent, catalyst loading, ratio of ligand(s) to metal, a base (which is often inorganic and has varying partial solubility in different solvents at different temperatures), and possible additives. Thus, the actual catalytic reactions can be very complex mixtures, which are challenging to describe with descriptors for AI/ML. The last challenge to overcome is that the majority of available data for catalytic reactions in the literature are reaction yields, which cannot be easily linked to $\Delta G^{\ddagger}$ due to interference from side reaction and unreliable reported reaction times, *i.e.* reactions are often left for a fixed time rather than monitored kinetically. These make the practical application of AI/ML chemical space exploration and prediction models for transition metal catalysed reactions uniquely challenging.

Nevertheless, a number of successful studies in the last three years have shown that some of these problems can be overcome with innovative approach and care while applying ML algorithms. The prevalent approach focuses on reaction optimisation, through predicting and optimising selec-

10.1002/cctc.202301475

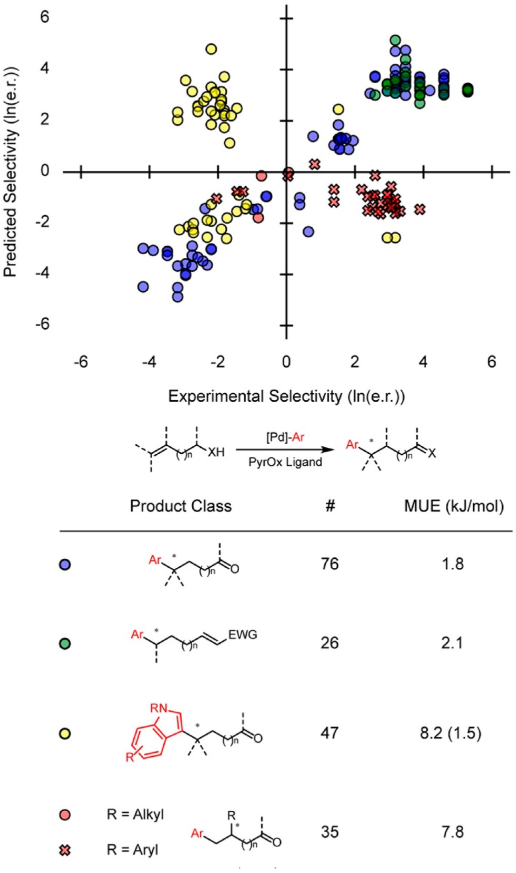


**Figure 8.** Comparison of 184 predicted and experimental selectivities. Stereochemistry (R/S) is indicated by +/- values, respectively. The MUE of the magnitudes of the selectivity, omitting absolute configuration, is included in parentheses. (reprinted with permission from ACS)[60]

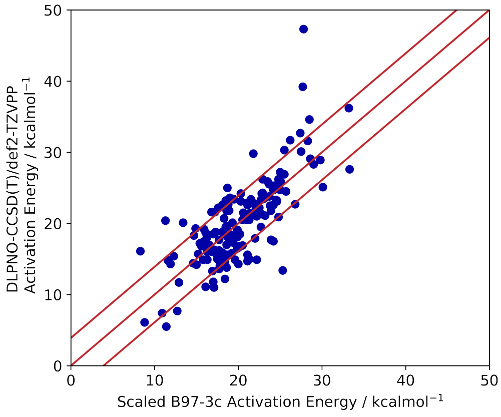


**Figure 9.** Scaled B97-3c activation energies of 68 TSOA and 83 TSSig transition states from ligands_lit_set, compared to their DLPNO–CCSD(T)/def2–TZVPP calculated activation energies. The red lines represent 3.9 kcal·mol$^{-1}$ (the MAE in the calculations) (reprinted with permission from RSC)[51]

tivity, particularly stereoselectivity. This has the advantage of avoiding the unreliable reaction yield and reaction time data. The $\Delta\Delta G^{\ddagger}$ value is linked to the two TS giving the two stereoisomers and the observed stereoselectivity under kinetic control. The $\Delta\Delta G^{\ddagger}$ value can be predicted with an appropriate regression ML algorithm, based on the amount of available data and the number of descriptors which can be fed into the algorithm. The main challenge is the high computational cost of calculating the TS with sufficient accuracy to predict enantioselectivity (a $\Delta\Delta G^{\ddagger}$ value of 2.1 kcal·mol$^{-1}$ would translate to a selectivity of >99% e.e.).

Thus, methods to approximate DFT results of TS with lower computational cost are essential. This was demonstrated by Wiest and Norrby in predicting stereoselectivity for the Pd-catalysed 1,4-conjugate addition of aryl boronic acids to enones.[75] A TSFF model was developed for the reaction, based on Q2MM calculations and MM3* force field and training TS data generated with M06/LANL2DZ/6-31+G*[76] The TSFF model was then incorporated into CatVS tool to carry out a conformational search.[59] Boltzmann averages for all of the conformers of the four different TS for each catalyst were used to calculate the enantiomeric ratio of the products against experimental values. The predictions were validated experimentally using an automated screen of 9 ligands, 38 aryl boronic acids, and 22 enones, leading to a mean unassigned error (MUE) of 1.8 kcal·mol$^{-1}$ and a R$^2$ value of 0.88 over 82 examples (Figure 10). This TSFF model was then used to carry out a virtual screen with 27 ligands and 59 enones. Selected results for 6-substituted pyrox ligands, which were not part of the training set, showed discrepancies against DFT calculations. Unfortunately, experimental validation was hampered by the synthesisability of the ligands. Relatively similar accuracy was previously observed when CatVS was applied to $OsO_4$-catalysed *cis*-dihydroxylation and Rh-catalysed asymmetric hydrogenation.[59]

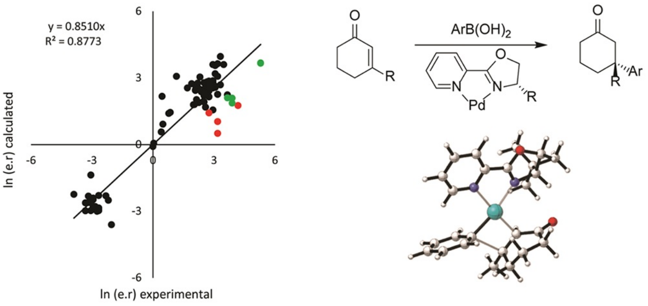


**Figure 10.** TSFF prediction of enantioselectivity in Pd-catalysed 1,4-conjugate addition of aryl boronic acids (reprinted with permission from ACS)[75]

An alternative approach is making predictions on $\Delta\Delta G^{\ddagger}$ based on the properties of the catalytic intermediates or starting materials, rather than those of the TS. This has the benefit of avoiding costly and error-prone DFT optimisation and frequency calculation of TS, which can account for >90% of the total CPU time of a campaign.[51] However, care must be taken to ensure that the generated regression models are robust in extrapolation beyond the training set. The standard practice of having a separate training set, test set and validation set is strongly recommended, although practical limitations often prevent it. Sigman and

10.1002/cctc.202301475

Toste has demonstrated the effectiveness of this approach, using the MO, vibrational and steric descriptors (generated with M06-2X/def2–TZVP//M06-2X/6-31+G(d,p)) of the chiral phosphoric acid counterion and the nucleophile to predict enantioselectivity in a Pd-catalysed intramolecular allylic substitution.[77] Multivariate Linear Regression (MLR) models were built with up to 35 descriptors for 16 experimental data points. These interpretable models led to mechanistic insights on the origin of enantioselectivity through multiple noncovalent interactions in this dual catalytic system. Another study was reported by Hong and Ackermann, used ML to design and optimise chiral carboxylic acids for cobalt-catalysed C-H alkylations.[78] While the $R^2$ and MAE values (using Linear Support Vector Regression algorithm) are not as good as those reported by Sigman and Toste, the catalyst, chiral carboxylic acids, reactants and reaction temperature (108 descriptors including buried volume, Sterimol, Fukui function, charge, bond dissociation energies, etc. for 59 reactions) were all included with 3D descriptors generated with *rdkit* (steric) and GFN2-xTB (charge, bond order and MOs). The models were used to predict carboxylic acids which give high enantioselectivity and yield, which were successfully validated experimentally. A very similar approach was used to predict enantioselectivity in a pallada–electrocatalysed C–H activation reaction based on 127 experimental datapoints.[79] A total of 119 descriptors were used, including 13 for the experimental conditions (*e.g.* properties of solvent, electrolyte, temperature and current), and ET algorithms was found to be the most effective, giving $R^2 = 0.91$ and MAE $= 0.236$ kcal·mol$^{-1}$.

An additional benefit of using descriptors based on the catalyst and starting materials is that some, if not all, the descriptors for different reactions can be reused. This is particularly true with ligand/metal combinations which have rigid structures around the metal regardless of the other ligands. One example is Cu-bisoxazoline (BOX) catalysts, which have been used as catalysts for enantioselective cyclopropanation, Diels-Alder cycloadditions, and difunctionalisation of alkenes. Sigman group showed that mechanism-specific categorisation of curated data sets and parameterisation of reaction components allow for simultaneous analysis of disparate organometallic intermediates such as carbenes and Lewis acid adducts.[15] Experimental data were curated from the literature on carbene, Lewis acid and radical-based transformations, i.e. 68 data points from 10 publications spanning a selectivity window of 0-99% *ee.* (0.0-3.1 kcal/mol). Comparison of ligand descriptors (M06-2X/def2-TZVP//B3LYPD3BJ/6-31(d,p)/LANL2DZ(Cu)) and their weighted contribution in each model reveals the relevant structural requirements necessary for high selectivity. The prediction errors ranged from $0.15 \pm 0.10$ to $0.79 \pm 0.42$ kcal·mol$^{-1}$ (most predictions within 10% *ee*), depending on the reaction (Figure 11). The scarcity of high quality experimental data is a key obstacle in applying AI/ML to catalysis, and this work showcased a possible workflow to combine experimental data for related ligand classes in catalytic reactions with similar stereo-inducing mechanistic steps. While the model for Cu catalysts cannot be directly applied to Fe/Ni/Mg/Pd-BOX catalysts, a separate unified MLR model was found to work on a new combined data set of 24 data points for these metals and showed similar accuracy.

Outside stereoselectivity, Sigman and Nozaki applied the methods described above to optimise phosphine–sulfonate ligands in Pd-catalysed copolymerisation of ethylene and methyl acrylate.[80] The data was filtered based on reaction temperature (80 °C for homopolymerisation, 80-100 °C for copolymerisation). A total of 62 descriptors were used with 112 experimental data points to predict the log(MW) of the polymer products, which depends more on the stability of the catalyst than on its activity. The models were built with PLS and LASSO algorithms, and led to the identification of ligand features which has high impacts on the MW of the product, such as the size of substituents on the phosphorus atom, the electron density and $d_{z^2}$ occupancy of the Pd atom, and the bite angle of the ligand. A MLR model for reaction yield of a Pd-Catalysed cyanation of aryl boronic acids based on mono– and diphosphine ligands was also reported based on the *kraken* dataset.[81] A wider substrate scope was accommodated, with tolerance for boronic acids bearing electron-withdrawing substituents.

On the other hand, Ess group reported an interesting study in which DFT-calculated TS and ML were combined to identify important features for selective olefin oligomerisation with Cr-catalysts.[82] A Random Forrest model was built to predict chemoselectivity in oligomerisation of ethylene into 1-octene *vs* 1-hexene ($\Delta\Delta G^{\ddagger}$) with RMSE $= 0.344$ kcal·mol$^{-1}$, based on 105 TS with Cr(P,N) catalysts and 14 molecular descriptors, i.e. bond lengths, angles, dihedrals, percent volume buried, and Cr metal center distance out pocket. Feature importance analysis of the model identified Cr–N distance, Cr–$\alpha$ distance (distance from Cr to an agostic C–H), and distance out of pocket (distance to the line between two ligating atoms of the ligand) as the most important features in enhancing selectivity for 1-octene, which informed subsequent catalyst designs.

In addition, Aspuru-Guzik, Hein and Sigman developed a closed-loop system to optimise a Suzuki-Miyaura coupling reaction on a vinyl tosylate substrate with stereoretention in batch (Figure 12a).[83] The process parameters included temperature, amount of boronic acid, palladium loading, ligand/palladium ratio, and the ligand as a discrete parameter. The Phoenics and Gryffin algorithms were employed to maximise the yield of the *E*-product and to minimise the palladium loading and aryl boronic acid equivalents.[84,85] Commercially available monodentate phosphines (365 ligands) were used to define the chemical space through a set of DFT descriptors which were subsequently condensed into 4 principal components. Subsequently, *k*-means clustering was carried out on these to divide the ligand chemical space into 24 regions, which were used to guide the ligand selection process for screening. A total of 192 experiments were required to optimise the outcome of the reaction, which compared well with the traditional approach using Design of Experiments (DoE), which will likely require $3^3 \times 23$ ligands $= 621$ experiment to achieve the same objectives.

Mack and Sigman demonstrated another multi-objective optimisation, *i.e.* for yield and selectivity, of two sequential catalytic reactions: an asymmetric Pd-catalysed Hayashi-Heck reaction and an asymmetric Rh-catalysed hydroformylation through optimisation of the biphosphine ligands (Figure 12b).[16] These reactions are the first two steps of the synthesis of a TRPA1 inhibitor and were optimised separately in this study.[86] In addition to the previously employed steric and electronic descriptors, quadrant-specific descriptors for ligands containing different symmetry elements were included, *e.g.* the percent buried volume ($V_{bur}$).[87,88] Application of a classification algorithm to the high through-

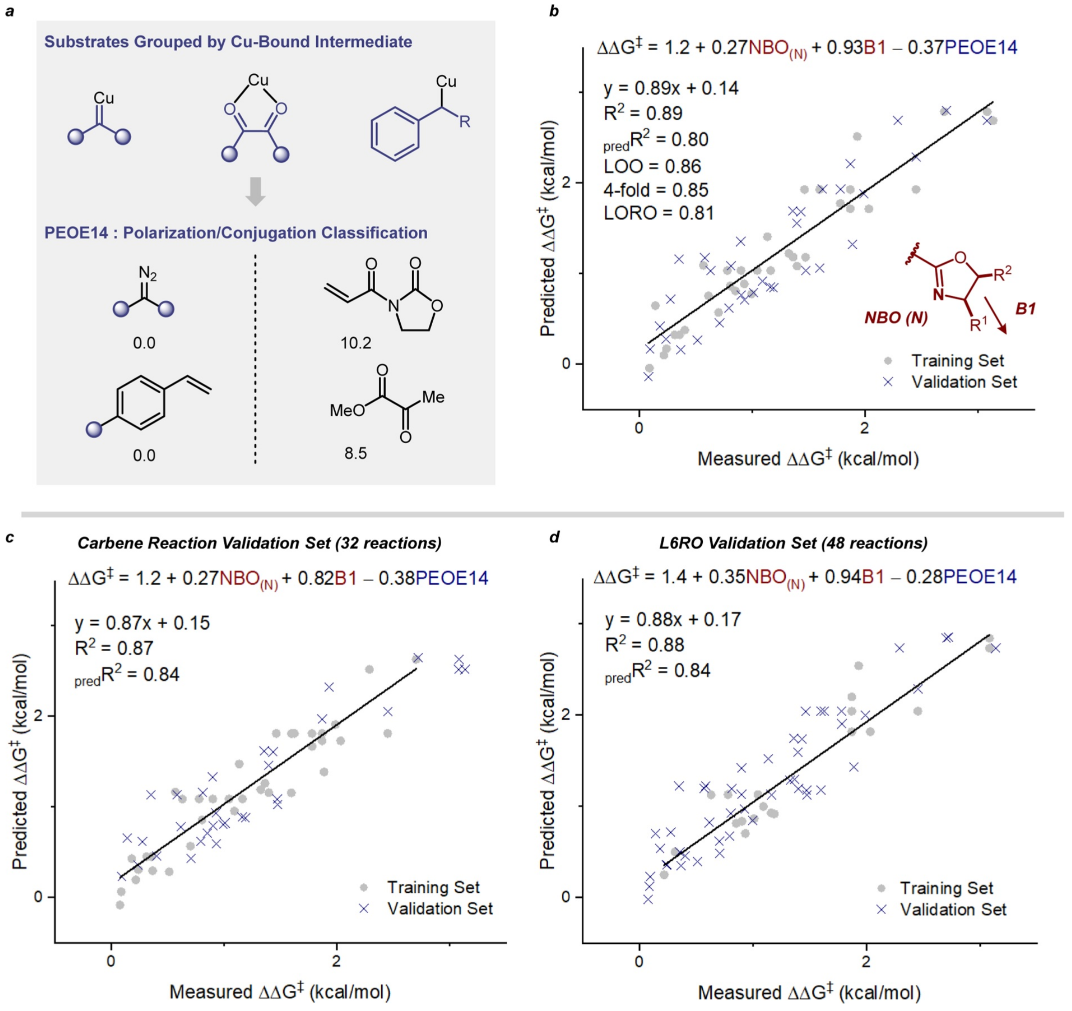


**Figure 11.** (a) Grouping of substrates based on proposed intermediates and subsequent classification using PEOE14 descriptor. (b) Multivariate regression analysis of Cu-BOX-catalyzed reactions (68 reactions). Plot of cross-validation [LOO and *k*-fold (k = 4)] and external validation ($_{pred}R^2$) by pseudorandom 50:50 partitioning of data into training set: validation set. (c) Plot of carbene-based reactions (32 reactions) being removed and held as a validation set. (d) Plot of six individual publications being removed (48 total reactions) and held as a validation set (L6RO = leave six reactions out). (Reprinted with permission from ACS)[15]

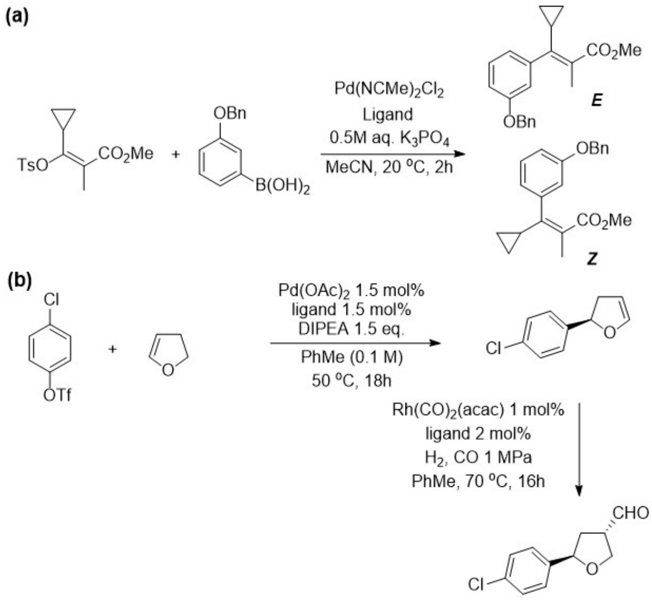


**Figure 12.** (a) Suzuki-Miyaura coupling reaction on a vinyl tosylate substrate for closed-loop optimisation;[83] (b) consecutive asymmetric Pd-catalysed Hayashi-Heck reaction and asymmetric Rh-catalysed hydroformylation for ligand optimisation.[16]

put experimental data of the Hayashi-Heck reaction strongly linked reactivity with the phosphorus lone pair occupancy of the ligand, in agreement with established mechanistic understanding.[89] Application of a logistic regression classifier on the hydroformylation step identified the buried volume and total ligand dipole as important descriptors. For regioselectivity of the first step, an MLR algorithm found strong correlation to two parameters, the computed anisotropic phosphorus NMR shielding and the occupancy of the $\sigma^*$ orbitals of the P–C bonds. After removal of some outliers, a two-term MLR model was found for hydroformylation regioselectivity using an electronic parameter PC occupancy and a steric parameter %$V_{bur}$NE. This $V_{bur}$ descriptor was also found to be linked to enantioselectivity. Taken together, the workflow led to the identification of the less often used ligands ($S$)-HexaMeO-BIPHEP and Walphos W003 for the two steps, which were validated experimentally to give the final product in excellent yield and purity.

## Conclusions and outlook

Application of AI and ML to transition metal catalysis is a rapidly evolving field of research with unique challenges. Recent publications have shown that it is possible to predict $\Delta G^\ddagger$ and $\Delta\Delta G^\ddagger$ with ML to reduce the resources required for direct TS calculations. They included a wide range of catalytic reactions, including coupling reactions, 1,4-addition of boronic acids to enones, cyclopropanation and oxidation of methane to methanol. Many of these successes focused on optimisation of the catalyst or catalytic process with relatively small experimental datasets. In this context, the applicable ML algorithms are somewhat limited, and the most advanced neural networks are often excluded. Thus, the scarcity of high quality experimental data, particularly kinetic data, is a key obstacle which needs to be addressed in order to progress the field.[90]

Chemical space exploration for transition metal catalysts has been effectively demonstrated using a combinatorial approach with monophosphines and porphyrin-type ligands. Wider chemical space exploration may be supported using the CSD, or drug-design cheminformatics techniques. Some recent publications have also reported guided exploration, instead of a randomised approach, of transition metal chemical space to achieve more efficient coverage.

Lastly, the complexity of transitional metal catalysis led to highly complex chemical systems which need to be addressed with reliable conformational searches, 3D structural and electronic descriptors and DFT methods which balance between computational cost and accuracy. These are unique challenges and are where exciting innovations and discoveries will be made in this area of research. In this regard, the work of Balcells group, including their recent publication on predicting organometallic properties with natural quantum graphs,[91] and the autonomous reaction network exploration in catalysis by Reiher indicate exciting developments in computational catalysis in the near future.[92]

## Acknowledgements

Samuel Mace thanks GoldenKeys High-Tech Materials Co., Ltd. (915-20900MA6DL1ER7N) and University of Leeds for his studentship. We thank Department of Science and Technology of Guizhou Province, Guizhou Service Center For Academician Expert, and Mr. Songlin Liu from WengFu Research Institute of Technology for their help in setting up this collaboration.

## Conflict of Interest

The authors declare no conflict of interests.

**Keywords:** Catalysis • Machine learning • Transition metal • Organometallics • Chemical space

## References

[1] A. V. Sadybekov, V. Katritch, *Nature* **2023**, *616*, 673.
[2] European Pharmaceutical Review, https://www.europeanpharmaceuticalreview.com/news/184106/first-ai-generated-small-molecule-drug-enters-phase-ii-tr accessed: 2023-10-04.
[3] G. Lefèvre, G. Franc, A. Tlili, C. Adamo, M. Taillefer, I. Ciofini, A. Jutand, *Organometallics* **2012**, *31*, 7694.
[4] J. W. Tye, Z. Weng, A. M. Johns, C. D. Incarvito, J. F. Hartwig, *Journal of the American Chemical Society* **2008**, *130*, 9971.
[5] G. O. Jones, P. Liu, K. N. Houk, S. L. Buchwald, *Journal of the American Chemical Society* **2010**, *132*, 6205.
[6] H. L. Aalten, G. van Koten, D. M. Grove, T. Kuilman, O. G. Piekstra, L. A. Hulshof, R. A. Sheldon, *Tetrahedron* **1989**, *45*, 5565.
[7] V. V. Litvak, U. S. M. Shein, *Z. Org. Khim.* **1974**, *10*, 2360.
[8] J. Lindley, *Tetrahedron* **1984**, *40*, 1433.
[9] H. Weingarten, *The Journal of Organic Chemistry* **1964**, *29*, 3624.

[10] C. Sambiagio, S. P. Marsden, A. J. Blacker, P. C. McGowan, *Chem. Soc. Rev.* **2014**, *43*, 3525.

[11] E. Sperotto, G. P. M. van Klink, G. van Koten, J. G. de Vries, *Dalton Trans.* **2010**, *39*, 10338.

[12] H.-Z. Yu, Y.-Y. Jiang, Y. Fu, L. Liu, *Journal of the American Chemical Society* **2010**, *132*, 18078.

[13] P.-F. Larsson, A. Correa, M. Carril, P.-O. Norrby, C. Bolm, *Angewandte Chemie International Edition* **2009**, *48*, 5691.

[14] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186.

[15] J. Werth, M. S. Sigman, *ACS Catalysis* **2021**, *11*, 3916.

[16] J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Puntener, K. A. Mack, M. S. Sigman, *Journal of the American Chemical Society* **2022**, *145*, 110.

[17] G. J. Sherborne, S. Adomeit, R. Menzel, J. Rabeah, A. Brückner, M. R. Fielding, C. E. Willans, B. N. Nguyen, *Chem. Sci.* **2017**, *8*, 7203.

[18] K. Jorner, A. Tomberg, C. Bauer, C. Sköld, P.-O. Norrby, *Nature Reviews Chemistry* **2021**, *5*, 240.

[19] A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves, H. J. Kulik, *Chemical Reviews* **2021**, *121*, 9927.

[20] L.-H. Mou, T. Han, P. E. S. Smith, E. Sharman, J. Jiang, *Advanced Science* **2023**, *10*, 2301020.

[21] V. von Burg, G. H. Low, T. Häner, D. S. Steiger, M. Reiher, M. Roetteler, M. Troyer, *Phys. Rev. Res.* **2021**, *3*, 033055.

[22] J. M. Crawford, C. Kingston, F. D. Toste, M. S. Sigman, *Accounts of Chemical Research* **2021**, *54*, 3136.

[23] S. Palkovits, *ChemCatChem* **2020**, *12*, 3995.

[24] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K. T. Schütt, J. Westermayr, M. Gastegger, R. J. Maurer, B. Kalita, K. Burke, R. Nagai, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Esters, D. Hicks, C. Toher, P. V. Balachandran, I. Tamblyn, S. Whitelam, C. Bellinger, L. M. Ghiringhelli, *Electronic Structure* **2022**, *4*, 023004.

[25] H. H. Cramer, S. Das, M. D. Wodrich, C. Corminboeuf, C. Werlé, W. Leitner, *Chem. Sci.* **2023**, *14*, 2799.

[26] M. D. Wodrich, A. Fabrizio, B. Meyer, C. Corminboeuf, *Chem. Sci.* **2020**, *11*, 12070.

[27] M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon, C. Corminboeuf, *ACS Catalysis* **2020**, *10*, 7021.

[28] D. J. Durand, N. Fey, *Accounts of Chemical Research* **2021**, *54*, 837.

[29] N. Fey, A. Koumi, A. V. Malkov, J. D. Moseley, B. N. Nguyen, S. N. G. Tyler, C. E. Willans, *Dalton Trans.* **2020**, *49*, 8169.

[30] D. J. Durand, N. Fey, *Chemical Reviews* **2019**, *119*, 6561.

[31] J. Jover, N. Fey, *Dalton Trans.* **2013**, *42*, 172.

[32] J. Jover, N. Fey, J. N. Harvey, G. C. Lloyd-Jones, A. G. Orpen, G. J. Owen-Smith, P. Murray, D. R. Hose, R. Osborne, M. Purdie, *Organometallics* **2012**, *31*, 5302.

[33] N. Fey, M. Garland, J. P. Hopewell, C. L. McMullin, S. Mastroianni, A. G. Orpen, P. G. Pringle, *Angewandte Chemie International Edition* **2012**, *51*, 118.

[34] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes **2022**.

[35] R. Gó mez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 268.

[36] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial Autoencoders **2016**.

[37] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **1986**, *323*, 533.

[38] A. Nandy, C. Duan, C. Goffinet, H. J. Kulik, *JACS Au* **2022**, *2*, 1200.

[39] N. Arunachalam, S. Gugler, M. G. Taylor, C. Duan, A. Nandy, J. P. Janet, R. Meyer, J. Oldenstaedt, D. B. K. Chu, H. J. Kulik, *The Journal of Chemical Physics* **2022**, *157*, 184112.

[40] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, *Journal of the American Chemical Society* **2022**, *144*, 1205.

[41] T. Gensch, S. R. Smith, T. J. Colacot, Y. N. Timsina, G. Xu, B. W. Glasspoole, M. S. Sigman, *ACS Catalysis* **2022**, *12*, 7773.

[42] J. M. Crawford, T. Gensch, M. S. Sigman, J. M. Elward, J. E. Steves, *Organic Process Research & Development* **2022**, *26*, 1115.

[43] J. P. Janet, C. Duan, T. Yang, A. Nandy, H. J. Kulik, *Chem. Sci.* **2019**, *10*, 7913.

[44] J. P. Janet, H. J. Kulik, *Chem. Sci.* **2017**, *8*, 5137.

[45] J. P. Janet, C. Duan, A. Nandy, F. Liu, H. J. Kulik, *Accounts of Chemical Research* **2021**, *54*, 532.

[46] J. P. Janet, H. J. Kulik, *The Journal of Physical Chemistry A* **2017**, *121*, 8939.

[47] A. Nandy, M. G. Taylor, H. J. Kulik, *The Journal of Physical Chemistry Letters* **2023**, *14*, 5798.

[48] C. Duan, S. Chen, M. G. Taylor, F. Liu, H. J. Kulik, *Chem. Sci.* **2021**, *12*, 13021.

[49] J. P. Perdew, K. Schmidt, *AIP Conference Proceedings* **2001**, *577*, 1.

[50] C. Duan, A. Nandy, G. G. Terrones, D. W. Kastner, H. J. Kulik, *JACS Au* **2023**, *3*, 391.

[51] M. A. S. Short, C. A. Tovee, C. E. Willans, B. N. Nguyen, *Catal. Sci. Technol.* **2023**, *13*, 2407.

[52] A. Nandy, C. Duan, J. P. Janet, S. Gugler, H. J. Kulik, *Industrial & Engineering Chemistry Research* **2018**, *57*, 13973.

[53] E. I. Ioannidis, T. Z. H. Gani, H. J. Kulik, *Journal of Computational Chemistry* **2016**, *37*, 2106.

[54] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016**, *72*, 171.

[55] O. Korb, B. Kuhn, J. Hert, N. Taylor, J. Cole, C. Groom, M. Stahl, *Journal of Medicinal Chemistry* **2016**, *59*, 4257.

[56] J. Halpern, *Science* **1982**, *217*, 401.

[57] P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, D. Balcells, *Chem. Sci.* **2020**, *11*, 4584.

[58] A. R. Rosales, T. R. Quinn, J. Wahlers, A. Tomberg, X. Zhang, P. Helquist, O. Wiest, P.-O. Norrby, *Chem. Commun.* **2018**, *54*, 8294.

[59] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows,

K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nature Catalysis* **2019**, *2*, 41.

[60] A. R. Rosales, S. P. Ross, P. Helquist, P.-O. Norrby, M. S. Sigman, O. Wiest, *Journal of the American Chemical Society* **2020**, *142*, 9700.

[61] M. Burai Patrascu, J. Pottel, S. Pinus, M. Bezanson, P.-O. Norrby, N. Moitessier, *Nature Catalysis* **2020**, *3*, 574.

[62] C. R. Corbeil, S. Thielges, J. A. Schwartzentruber, N. Moitessier, *Angewandte Chemie International Edition* **2008**, *47*, 2635.

[63] N. Weill, C. R. Corbeil, J. W. De Schutter, N. Moitessier, *Journal of Computational Chemistry* **2011**, *32*, 2878.

[64] C. Corbeil, S. Thielges, J. Schwartzentruber, N. Moitessier, *Angewandte Chemie-German Edition* **2008**, *120*, 2675.

[65] K. Jorner, T. Brinck, P.-O. Norrby, D. Buttar, *Chemical Science* **2021**, *12*, 1163.

[66] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, K. Burke, *Nature communications* **2020**, *11*, 5223.

[67] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, *Science Advances* **2017**, *3*, e1603015.

[68] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, A. Tkatchenko, *Computer Physics Communications* **2019**, *240*, 38.

[69] S. Chmiela, H. E. Sauceda, K.-R. Müller, A. Tkatchenko, *Nature Communications* **2018**, *9*, 3887.

[70] M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, S. Grimme, *Phys. Chem. Chem. Phys.* **2021**, *23*, 287.

[71] H. Neugebauer, B. Bädorf, S. Ehlert, A. Hansen, S. Grimme, *Journal of Computational Chemistry* **2023**, *44*, 2120.

[72] S. Chen, T. Nielson, E. Zalit, B. B. Skjelstad, B. Borough, W. J. Hirschi, S. Yu, D. Balcells, D. H. Ess, *Topics in Catalysis* **2022**, *65*, 312.

[73] C. Duan, J. P. Janet, F. Liu, A. Nandy, H. J. Kulik, *Journal of Chemical Theory and Computation* **2019**, *15*, 2331.

[74] C. Duan, A. Nandy, H. Adamji, Y. Roman-Leshkov, H. J. Kulik, *Journal of Chemical Theory and Computation* **2022**, *18*, 4282.

[75] J. Wahlers, M. Maloney, F. Salahi, A. R. Rosales, P. Helquist, P.-O. Norrby, O. Wiest, *The Journal of Organic Chemistry* **2021**, *86*, 5660.

[76] N. L. Allinger, Y. H. Yuh, J. H. Lii, *Journal of the American Chemical Society* **1989**, *111*, 8551.

[77] C.-C. Tsai, C. Sandford, T. Wu, B. Chen, M. S. Sigman, F. D. Toste, *Angewandte Chemie International Edition* **2020**, *59*, 14647.

[78] Z.-J. Zhang, S.-W. Li, J. C. A. Oliveira, Y. Li, X. Chen, S.-Q. Zhang, L.-C. Xu, T. Rogge, X. Hong, L. Ackermann, *Nature Communications* **2023**, *14*, 3149.

[79] L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. A. Oliveira, S.-W. Li, L. Ackermann, X. Hong, *Nature Synthesis* **2023**, *2*, 321.

[80] S. Akita, J.-Y. Guo, F. W. Seidel, M. S. Sigman, K. Nozaki, *Organometallics* **2022**, *41*, 3185.

[81] J. De Jesus Silva, N. Bartalucci, B. Jelier, S. Grosslight, T. Gensch, C. Schünemann, B. Müller, P. C. J. Kamer, C. Copéret, M. S. Sigman, A. Togni, *Helvetica Chimica Acta* **2021**, *104*, e2100200.

[82] S. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chem. Sci.* **2020**, *11*, 9665.

[83] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Communications Chemistry* **2021**, *4*, 112.

[84] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 1134.

[85] F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, A. Aspuru-Guzik, *Applied Physics Reviews* **2021**, *8*, 031406.

[86] J. A. Terrett, H. Chen, D. G. Shore, E. Villemure, R. Larouche-Gauthier, M. Déry, F. Beaumier, L. Constantineau-Forget, C. Grand-Maître, L. Lépissier, S. Ciblat, C. Sturino, Y. Chen, B. Hu, A. Lu, Y. Wang, A. P. Cridland, S. I. Ward, D. H. Hackos, R. M. Reese, S. D. Shields, J. Chen, A. Balestrini, L. Riol-Blanco, W. P. Lee, J. Liu, E. Suto, X. Wu, J. Zhang, J. Q. Ly, H. La, K. Johnson, M. Baumgardner, K.-J. Chou, A. Rohou, L. Rougé, B. S. Safina, S. Magnuson, M. Volgraf, *Journal of Medicinal Chemistry* **2021**, *64*, 3843.

[87] L. Falivene, R. Credendino, A. Poater, A. Petta, L. Serra, R. Oliva, V. Scarano, L. Cavallo, *Organometallics* **2016**, *35*, 2286.

[88] A. Poater, B. Cosenza, A. Correa, S. Giudice, F. Ragone, V. Scarano, L. Cavallo, *European Journal of Inorganic Chemistry* **2009**, *2009*, 1759.

[89] D. Mc Cartney, P. J. Guiry, *Chem. Soc. Rev.* **2011**, *40*, 5122.

[90] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, *Journal of the American Chemical Society* **2021**, *143*, 18820.

[91] H. Kneiding, R. Lukin, L. Lang, S. Reine, T. B. Pedersen, R. De Bin, D. Balcells, *Digital Discovery* **2023**, *2*, 618.

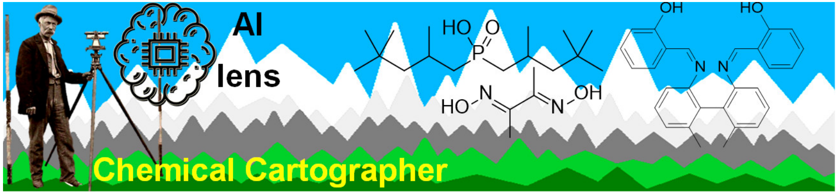
[92] M. Steiner, M. Reiher, *Topics in Catalysis* **2022**, *65*, 6.

10.1002/cctc.202301475

## Entry for the Table of Contents

TOC Graphic
option 1
max. 5.5 x 5.0 cm

Authors should provide a short Table of Contents graphical abstract and accompanying text (up to 450 characters including spaces). The graphical abstract should stimulate curiosity. Repetition or paraphrasing of the title and experimental details should be avoided.



A critical review of the use of Artificial Intelligence and Machine Learning to discover and optimise organometallic catalysts. Topics include methods for exploring transition metal chemical space, generating descriptors from molecular modelling, and their applications to complex catalytic systems.