ChemComm

HIGHLIGHT

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2025, **61**, 18247

Catalysis meets machine learning: a guide to data-driven discovery and design

Eleonora Casillo, [a] Thomas Scattolin *[b] and Steven P. Nolan *[a]

Received 12th September 2025,
Accepted 23rd October 2025

DOI: 10.1039/d5cc05274b

rsc.li/chemcomm

Machine learning (ML) has rapidly become an indispensable tool across the chemical sciences, offering powerful methods to extract patterns from data and make accurate predictions in complex, multi-dimensional systems. In organometallic catalysis, its potential is particularly evident: while transition-metal catalysed reactions are at the core of modern synthesis, their design and optimization remain challenging due to the vastness of chemical space, the scarcity of standardized data, and the intricate interplay of steric, electronic, and mechanistic factors. This review aims to provide chemists with both a conceptual and practical entrypoint into the field, beginning with a concise introduction to the principles of ML and its most widely used algorithms. It then surveys recent advances by structuring the discussion according to applications: optimization of reaction conditions, mechanistic elucidation, ligand classification and design, stereocontrol, and the discovery of novel catalysts. By combining methodological insights with case studies, we highlight how ML can reduce experimental workload, enhance mechanistic understanding, and guide rational catalyst development, while also outlining current limitations and future opportunities at the interface of data science and catalysis.

1. Introduction

The optimization and study of chemical reactions traditionally rely on empirical methods, where chemists adjust parameters based on their understanding to achieve optimal outcomes. This approach is time-consuming and resource-intensive, most often relying on trial-and-error experimentation.[1] A central challenge in this process is the efficient identification of

[a] *Department of Chemistry and Center for Sustainable Chemistry, Ghent University, Krijgslaan 281 (S-3), 9000 Ghent, Belgium. E-mail: steven.nolan@ugent.be*
[b] *Dipartimento di Scienze Chimiche, Università degli Studi di Padova, via Marzolo 1, 35131 Padova, Italy. E-mail: thomas.scattolin@unipd.it*

Eleonora Casillo

Eleonora Casillo earned her MSc in Chemistry from Ca' Foscari University of Venice in 2023, conducting thesis research at the Department of Chemistry at Ghent University, under the supervision of Dr Thomas Scattolin and Prof. Steven P. Nolan. Her MSc research focused on the catalytic activity of platinum(II)-N-heterocyclic carbene and thioether complexes, as well as the application of machine learning. She is currently pursuing a PhD in the group of Prof. Nolan at Ghent University, where her research centres on the synthesis and application of transition metal-based complexes in both homogeneous and heterogeneous catalysis.

Thomas Scattolin

Thomas Scattolin completed his PhD in Chemistry in 2019 under the supervision of Prof. Fabiano Visentin at the University of Trieste. In 2019 he was a visiting scientist in the laboratories of Prof. Antonio Togni at the ETH in Zurich. In 2020, he joined the group of Prof. Steven P. Nolan at Ghent University as a postdoc researcher. Since 2022, he has been assistant professor in Inorganic Chemistry at the University of Padova. His research activity primarily focuses on the synthesis and reactivity of late transition metal complexes with applications in homogeneous catalysis and medicinal chemistry.

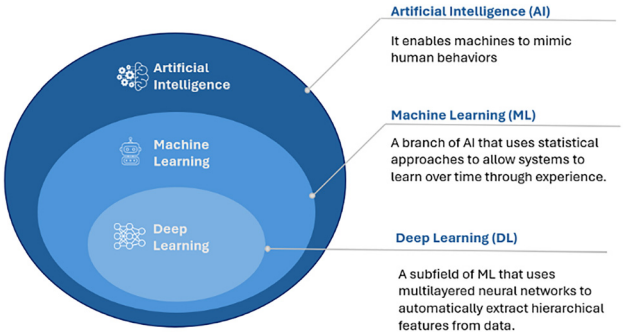

Fig. 1 Representation of artificial intelligence, machine learning and deep learning.

optimal reaction conditions within a limited experimental budget. This task is especially difficult in large and multi-dimensional reaction spaces, where time and cost constraints severely restrict experimental scope. Conventional strategies, often dependent on initial guesses and existing system knowledge, can be inefficient and expensive, particularly when starting points are distant from the ideal solution. Machine learning (ML) excels at extracting implicit knowledge from data by inferring functional relationships statistically, even without detailed problem-specific knowledge.[2] Unlike traditional approaches, ML starts from a generalized model and iteratively refines it, enabling early and efficient exploration of complex problems.[3] By combining data-driven algorithms with scientific theories, this interdisciplinary approach enhances the synergy between empirical data and theoretical frameworks, providing a powerful tool to navigate vast chemical spaces and accelerate the optimization process, all while deepening our understanding of complex catalytic systems.

Artificial Intelligence (AI) is becoming an essential tool in various branches of chemistry, being used to predict molecular properties, speed up computational simulations, design novel compounds, and propose viable synthetic pathways to new products.[4] Before surveying applications in organometallic catalysis, it is helpful to distinguish three terms that are often mingled (see Fig. 1):

- Artificial intelligence (AI): any computational method that performs tasks associated with human intelligence (reasoning, decision-making, language).[5]
- Machine learning (ML): algorithms that learn patterns from data to make predictions or decisions without hard-coded rules.[2]
- Deep learning (DL): a subfield of ML that uses multi-layer neural networks to model complex, nonlinear relationships; particularly effective with large, diverse datasets.[6]

2. The role and fundamentals of machine learning techniques in organometallic catalysis

The integration of ML into organometallic catalysis represents one of the most transformative trends of the past decade in chemical research. Transition metal-catalysed reactions are pillars of modern synthesis, yet their design and optimization remain labour-intensive: the design and optimization of such reactions often remain empirical, involving time-consuming and costly experimental trials. Traditional computational tools, such as density functional theory (DFT), offer mechanistic insight but are limited by their computational expense, particularly when navigating vast chemical spaces. Moreover, catalyst development is challenged by the complexity of reaction mechanisms, the high dimensionality of tuneable parameters, and the limited transferability of successful designs across different systems. The scarcity of standardized, high-quality experimental data and the difficulty of integrating molecular-scale insights with macroscopic performance further hinder progress. In recent years, ML has emerged as a powerful complement to both empirical and theoretical approaches.[2] By learning patterns from experimental or computed data, ML models can make accurate predictions about reaction yields, selectivity, optimal conditions, and even mechanistic pathways.[7]

The present contribution presents diverse applications of ML in organometallic catalysis over the last decade, organizing key developments by the nature of the application: optimization of reaction conditions, prediction of catalytic activity and enantioselectivity, ligand design, mechanism elucidation, and new catalyst discovery. Our aim is to provide a concise, yet thorough overview accessible to chemists across disciplines, emphasizing the field's growing impact and future directions.

In general terms, the foundation of ML rests on two critical components: data and algorithms.[8] Data refers to any type of input a computer can interpret (text, images, or sound). An algorithm is a defined sequence of steps the computer follows to analyze these data and learn from them. The combination of data and algorithms during the training process results in what is known as a ML model.

Steven P. Nolan

Steven P. Nolan received his PhD from the University of Miami where he worked under the supervision of Prof. Carl Hoff. After a postdoctoral stay with Prof. Tobin J. Marks at Northwestern University, he joined the University of New Orleans in 1990. In 2006 he joined the Institute of Chemical Research of Catalonia (ICIQ). In early 2009, he joined the School of Chemistry at the University of St Andrews and in 2017 joined the Department of Chemistry of Ghent University as Senior Full Professor. Professor Nolan's research interests revolve around the design and synthesis of catalytic complexes enabling organic transformations.

Table 1 Applications, advantages and limitations of the two key-methods: supervised and unsupervised learning

Aspect	Supervised learning	Unsupervised learning
Data required	Labeled	Unlabeled
Applications	Classification, regression	Clustering, association, dimensionality reduction
Advantages	High accuracy, interpretable results	Reveals hidden patterns, no need for labeled data
Limitations	Requires labeled data, time & money consuming	Lower predictive power, harder to interpret

There are three main learning paradigms: supervised, unsupervised, and hybrid (Table 1).[9]

Supervised learning learns a mapping from inputs to a labelled output (*e.g.*, predicting yield or enantioselectivity from ligand descriptors). It excels when labels are reliable and reasonably abundant.

Unsupervised learning finds structure in unlabelled data (*e.g.*, clustering ligands by descriptor similarity; dimensionality reduction to visualize reaction spaces). It is useful for hypothesis generation and dataset curation.

Hybrid/semi-supervised learning combines both (*e.g.*, pre-training on unlabelled structures and fine-tuning on a smaller labelled set) to improve data efficiency.

The core distinction between the supervised and unsupervised learning lies in the nature of the data used during training. To put it simply: supervised learning operates by training a model on a labelled dataset, where each input is paired with the correct output. This is analogous to teaching with a predefined curriculum: the algorithm is presented with known examples (*e.g.*, reactions with experimentally determined enantiomeric excess, % ee) and learns to map structural or mechanistic features to the target property. Once trained, the model can predict outcomes for new, unseen substrates or catalysts. In contrast, unsupervised learning identifies inherent patterns, groupings, or correlations within data without pre-existing labels. Here, the algorithm explores the dataset autonomously to discover latent structure, for instance, clustering catalysts or ligands based on similarity in their molecular descriptors or reaction outcomes.[9] In catalysis, this can reveal novel classifications of ligands that confer similar selectivity or activity, even in the absence of *a priori* mechanistic hypotheses. While it offers the advantage of discovering hidden structures without prior labeling, its results are often more difficult to interpret and generally less accurate when used for predictive purposes.[10] Hybrid learning integrates both supervised and unsupervised learning: a portion of the weights is typically determined through supervised learning, while the remaining weights are derived through unsupervised learning.[11]

The following section provides a concise overview of the most used ML approaches in catalysis, with the goal of arming readers with the essential background to interpret and appreciate the methods described throughout the next chapters.

3. Key machine learning algorithms for chemical applications

Several ML algorithms have proven particularly useful in chemical applications:

(i) Linear regression: one of the simplest models, linear regression assumes a direct, proportional relationship between descriptors and outcomes.[12] While often limited in complex systems, it serves as a baseline and is sometimes surprisingly effective in well-behaved chemical space. For example, Liu *et al.* used in 2022 a Multiple Linear Regression (MLR) to predict activation energies for C–O bond cleavage in Pd-catalyzed allylation.[12] Using DFT-calculated data from 393 reactions, they modeled energy barriers using different key descriptors. The resulting model (R^2 = 0.93) successfully captured electronic, steric, and hydrogen-bonding effects, demonstrating the MLR ability to quantify complex catalytic interactions across diverse chemical space.

(ii) Random Forest: a type of ensemble model composed of many decision trees (Fig. 3). Each tree is trained on a random subset of data, and the final prediction is an average (regression) or a vote (classification). A hypothetical challenge that a chemist might face is predicting whether a given organometallic complex will catalyse a reaction in high yield. Random Forest can take as input hundreds of molecular descriptors (*e.g.*, electronic properties, steric factors, geometries, orbitals *etc.*...) and learn a general rule by combining the decisions of multiple trees, each of which processes a different subset of data.[13] This approach allows to bring to light non-linear relationships within the data, leading to more accurate and robust predictions. An application in catalysis is reported by Doyle and colleagues who used a random forest model to predict reaction yields and screen reaction conditions in Pd-catalysed aminations and Suzuki–Miyaura couplings.[14]

(iii) Deep learning and artificial neural networks (ANNs): deep learning (DL) utilizes artificial neural networks (ANNs), with multiple hidden layers, to model complex, non-linear relationships within data.[15] These deep architectures are especially effective when applied to large and diverse datasets, enabling the automated extraction of hierarchical features. In chemistry, DL has been used to optimize multi-parameter reaction conditions and predict enantioselectivity with high accuracy.[16–18] The design of ANNs is conceptually inspired by the structure and function of biological neurons, which are the fundamental units of the nervous system. Biological neurons communicate through electrochemical signals and are composed of dendrites that receive input, a soma that is the cell body that process information, and an axon that transmits output to other neurons. Artificial neurons are simplified mathematical abstractions of this biological model (Fig. 2).[11]

ANNs are composed of basic units called nodes or processing elements and the connections between them. Each node receives inputs, either from other nodes or the external

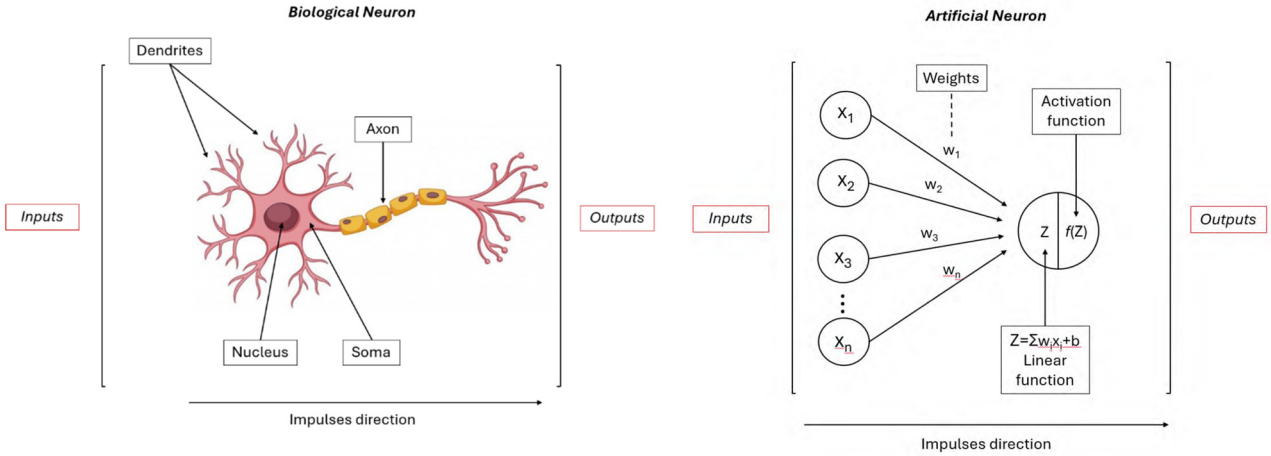

Fig. 2 Comparison between a biological neuron (left) and an artificial neuron (right).

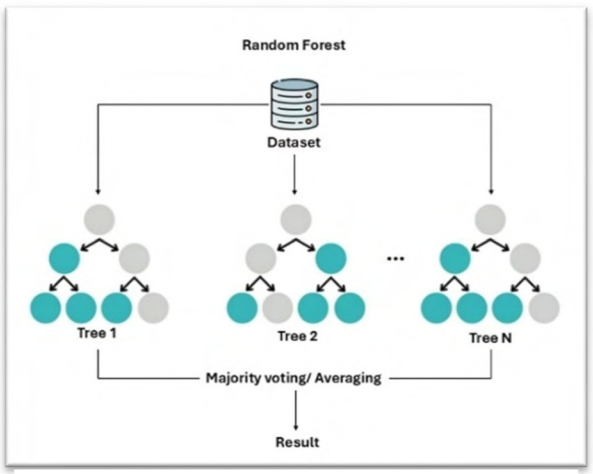

Fig. 3 Representation of random Forest model.

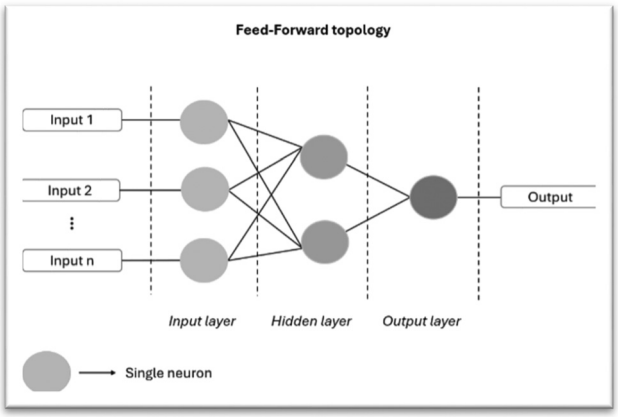

Fig. 4 Representation of feed forward topology.

environment, and produces outputs that influence others or interact with the environment.[11] A node processes its overall input through a specific function *f*, generating an output. Connections between nodes vary in strength, representing either excitation (positive values) or inhibition (negative values).[19] These connections can change over time, initiating a learning process across the network.[20] This process of adjustment is governed by what is known as the ''Law of Learning''.[21] Since these adjustments occur over time, learning in ANNs is inherently dynamic and driven by repeated interactions with the environment, represented by data.[22] Through this, ANNs can ''interpret'' their environment and its underlying relationships.[23–25] Neurons in an ANN can be arranged in various topologies, such as one- or two-dimensional layers, three-dimensional grids, or higher-dimensional structures, depending on the complexity and volume of input data.[11] The most widely used structure is the feed forward topology, where information always moves in one direction; it never goes backwards (Fig. 4).[26,27]

In this system a set of nodes forms the input layer, typically reflecting the number of input variables. Information then moves through one or more hidden layers before reaching the output layer, which delivers the result. While the resemblance is largely metaphorical, the learning mechanism in ANNs, typically based on backpropagation and gradient descent, bears conceptual similarity to synaptic plasticity in biological systems, where connection strengths are updated in response to experience or error feedback.[11] The backpropagation architecture is the most commonly used learning algorithm for training feed forward neural networks.[28] In this process, the network progressively identifies patterns and relationships within the data, updating its weights *via* backpropagation to reduce the discrepancy between predicted and true outputs. Even in their simpler forms, ANNs can capture significant non-linear patterns. When trained in chemical reaction data, such models have been used to predict optimal catalysts, reagents, and reaction conditions across diverse reaction classes, offering valuable support in reaction optimization and discovery.[29–33]

(iv) Graph neural networks: one of the most powerful and rapidly evolving classes of ML models, particularly in the domains of chemistry and materials science, is the graph neural networks (GNNs).[34,35] The representation of molecules and materials as graphs has deep historical roots in mathematical chemistry, dating back to the 19th century-preceding even the formal development of graph theory.[4] The GNNs defining capability lies in the natural way they operate on graph-structured data, making them ideally suited for modeling atomic and molecular systems.[36] While molecular graphs are typically undirected with well-defined atom and bond types, the extension of this concept to solid-state materials requires additional considerations. In crystalline systems, bonding is not always clearly defined, and the three-dimensional arrangement of atoms plays a more prominent role. Nonetheless, graph-based representations remain highly effective in capturing local environments and long-range structural information.

A core strength of GNNs is their ability to incorporate four essential types of information:

1. Node-level features (*e.g.*, atom types)
2. Edge-level features (*e.g.*, bond orders or distances)
3. Global context (*e.g.*, temperature, pressure)
4. Overall connectivity or topology of the graph.

In GNNs, molecules can be represented as graphs: in this representation, nodes correspond to atoms or atomic sites, while edges represent chemical bonds or spatial proximity between atoms, allowing GNNs to capture the underlying structure of chemical compounds and materials with high fidelity and process these graphs directly, learning to extract chemical information from structure alone.[37]

Fig. 5 illustrates how GNNs were used by Friederich and colleagues in chemistry and how their performance has been benchmarked.[4] Panel (a) shows a schematic of the message-passing principle: molecules or crystalline materials are represented as graphs, where atoms are nodes and bonds or spatial neighbours are edges. Through iterative message passing, each atom updates its representation by aggregating information from its neighbours, gradually capturing both local and long-range structural features. This enables the model to learn chemically meaningful descriptors directly from structure. Panel (b) reports the accuracy of different GNN architectures on the widely used QM9 dataset. QM9 is a collection of *ca.* 134 000 small organic molecules (up to nine heavy atoms: C, N, O, F, plus hydrogens), each optimized at the DFT level. For each molecule, more than a dozen physicochemical properties were computed, including 3D geometries, formation energies, free energies, dipole moments, polarizabilities, enthalpies of formation, vibrational frequencies, and frontier orbital levels (HOMO/LUMO). The graph shows mean absolute errors for predictions of total energy (red circles), HOMO (orange triangles), and LUMO (blue inverted triangles), highlighting the rapid improvement in predictive accuracy achieved by successive generations of GNNs.

Unlike conventional ML models that rely on manually crafted descriptors, GNNs autonomously learn internal feature representations through a process of message passing between neighboring nodes, followed by aggregation and readout steps that produce graph-level outputs. These models remove the need for explicit feature engineering and offer a more generalizable approach. GNNs are also capable of modeling complex

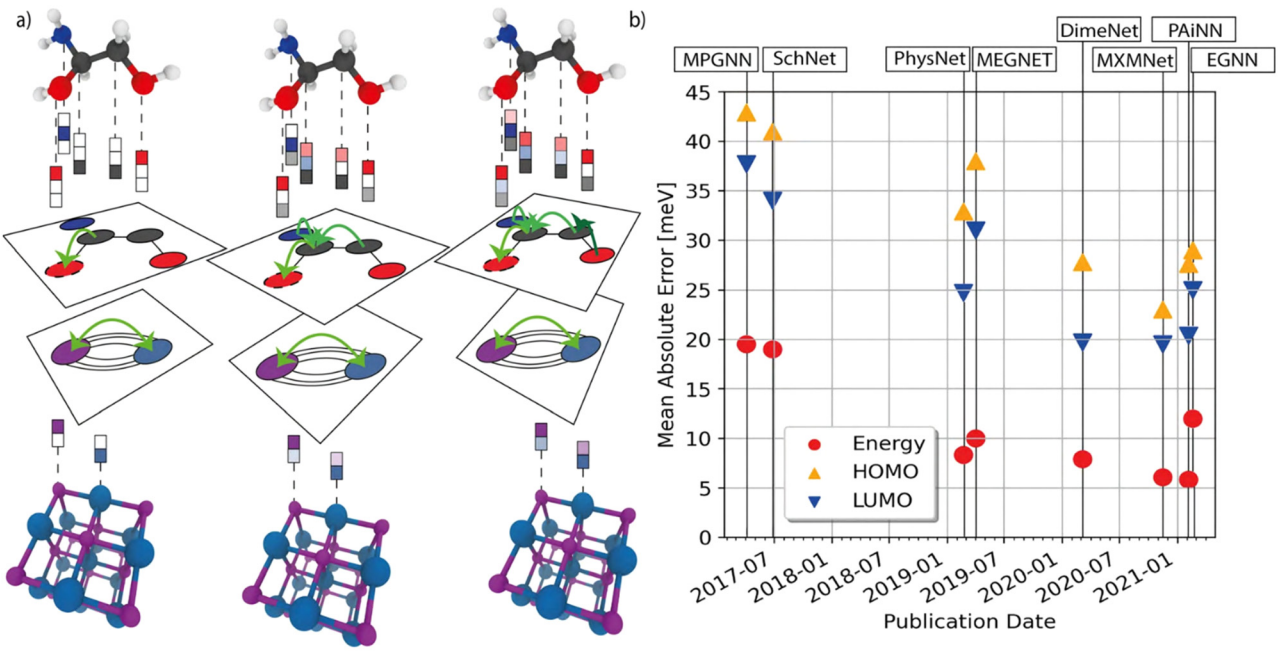

Fig. 5 (a) Schematic depiction of the message passing operation for molecules and crystalline materials. (b) QM9 benchmark. Mean absolute error of the prediction of internal (red circles), highest occupied molecular orbital (HOMO, orange triangles), and lowest unoccupied molecular orbital (LUMO, inverted blue triangles) energies for different GNN models since 2017. Reproduced from ref. 4 with permission from the communications materials, copyright 2022.

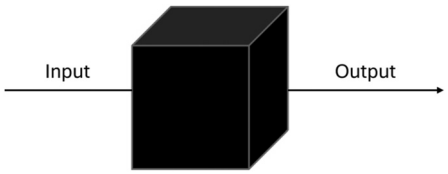

Fig. 6 Black box approach.

systems involving non-covalent interactions, doping, or structural disorder-scenarios where classical descriptors often fall short.[4] Given their capacity to learn directly from raw structural data and model intricate atomistic interactions, GNNs are poised to become a cornerstone of next-generation computational chemistry and materials design workflows.

(v) The ''black box'' challenge and Bayesian optimization: many ML models, especially neural networks and ensemble methods, are considered ''black boxes'' (Fig. 6). A black-box function is a system whose internal workings are not explicitly known, but whose outputs can be observed for given inputs.[38] In chemistry, the effect of a specific combination of temperature, catalyst, and solvent on reaction yield may not be precisely understood. Nevertheless, the experiment can be carried out and the outcome, such as yield, conversion, or enantioselectivity, empirically determined. This defines a black-box setting: the internal mechanism is opaque, but input–output behavior is measurable. Consequently, these techniques provide accurate predictions but offer limited insight into how decisions are made.[38] This lack of transparency can be problematic in scientific research, where mechanistic understanding is often crucial.

To address this, interpretability tools have been developed. One such tool is SHAP (SHapley Additive exPlanations),[39] which assigns important scores to each feature, showing how it contributed to a particular prediction.[40] While tools like SHAP improve interpretability by explaining individual predictions, they do not directly address how to efficiently search through large and complex input spaces to improve outcomes. In experimental sciences such as chemistry, interpretability may support mechanistic hypotheses, but practical constraints, such as time, material cost, and safety, often make it unfeasible to explore the full design space exhaustively.[40] This is where complementary strategies like Bayesian Optimization (BO) become essential. ML in the form of BO proves particularly suited to the challenge of chemical reaction optimization because it works with small datasets and can explore very large reaction spaces.[41] BO is not a predictive model *per se*, but rather a strategy for the global optimization of expensive black-box functions.[42] The term expensive does not necessarily refer to monetary cost, instead, it encompasses the overall experimental burden, including time, reagents, instrumentation, and labor. Because of this, the primary goal is often to minimize the number of experimental trials needed to achieve optimal results. BO addresses this by using a probabilistic surrogate model, typically a Gaussian Process, to estimate the relationship between inputs and the objective function.[42] It then applies an acquisition function to select the next experiment, balancing exploration, which prioritizes regions of the domain where the model uncertainty is high (it allows to discover regions that may contain better values but have not yet been sufficiently sampled), and exploitation which favours regions where the model predicts high objective values based on existing evidence.[43] This iterative, data-efficient process enables rapid identification of optimal reaction conditions or molecular features, even in high-dimensional and costly experimental spaces. BO can be broadly categorized into single-objective (SOBO) and multi-objective Bayesian optimization learning (MOBO).[44] In the single-objective setting, the goal is to optimize a single property of interest, such as catalytic activity or stability. It is used to optimize a single objective function, denoted $f(x)$. The goal is to identify the global optimum (maximum or minimum) of $f(x)$ over the defined search space. The output is typically a single best point x^*, based on the highest predicted utility from the surrogate model.[44]

The workflow begins with a starting dataset comprising candidate molecules and their experimentally measured or computed property values. A surrogate model, often a Gaussian Process, is trained on this dataset to approximate the objective function. Using this model, the properties of new and untested candidates drawn from a larger chemical library are predicted, along with associated uncertainties. An acquisition function then selects the most promising candidates based on the trade-off between exploration and exploitation.[43] These acquired candidates are subsequently evaluated (experimentally or *via* computational calculations), and their scores are fed back into the dataset to refine the surrogate model. In MOBO, that is an extension of the SOBO framework, the aim is to optimize several, often conflicting, objectives at the same time, such as catalytic efficiency and selectivity. Instead of a single objective, one considers a vector of objective functions.

The workflow follows a similar structure of the SOBO: a surrogate model is trained using a dataset containing molecular candidates and their multiple property values. However, instead of a single acquisition function, the selection step must consider the set of candidates that represent the best trade-offs among the objectives. The newly acquired candidates are then evaluated and used to update the dataset, refining the model.

Both approaches rely on probabilistic modeling to guide decision-making and reduce the experimental workload, rendering them especially well-suited to the complex, high-dimensional design landscapes characteristic of organometallic catalysis.

4. Molecular descriptors in machine learning for catalysis

ML models inherently operate on numerical data, necessitating the transformation of complex chemical entities, such as ligands, substrates, and catalysts, into quantitative formats for use in predictive tasks within catalysis. These numerical representations, referred to as molecular descriptors, encode

structural, electronic, steric, and topological characteristics of molecules. The selection and design of suitable descriptors constitute a critical step in the ML workflow, often exerting a decisive influence on model performance and predictive accuracy. Molecular descriptors are defined as numerical values that encapsulate specific properties of a molecule. Steric descriptors, such as the percent buried volume (% V), are commonly employed to quantify the spatial footprint of ligands around a metal center.[45] These descriptors are instrumental in assessing steric hindrance and catalyst accessibility, both of which are key factors in determining catalytic efficiency. Electronic descriptors, including quantities such as the HOMO–LUMO energy gap and dipole moments, characterize the electronic structure of molecules. They provide insights into reactivity, electrophilicity, and nucleophilicity, all of which are essential for modelling chemical behaviour. Geometric descriptors, for example bite angles and torsional angles, convey information about the three-dimensional conformation of molecules, which can critically influence the positioning of reactive sites and, consequently, the outcome of catalytic processes.

Lastly, topological descriptors, such as molecular fingerprints, encode information about molecular connectivity and atom-wise relationships in a format that does not require explicit geometric coordinates. These descriptors effectively transform molecular input into a format amenable to numerical analysis, enabling ML models to identify patterns and correlations that underlie catalytic behavior, reactivity, or selectivity. The process of feature engineering involves careful selection, calculation, and preprocessing of molecular descriptors prior to training the ML model. This step is often critical, as irrelevant or poorly chosen features can obscure meaningful patterns, degrade model performance, and increase the risk of overfitting.

First, it requires the identification of descriptors that are chemically meaningful and relevant to the specific problem under investigation. Next, the selected descriptors must be computed using a range of methods, which may include quantum chemical calculations (such as density functional theory, DFT), cheminformatics toolkits (such as RDKit), or semi-empirical approaches. Finally, the resulting descriptor matrix must be cleaned and pre-processed. This includes tasks such as removing redundant or highly correlated features, imputing missing values if present, and applying normalization or scaling procedures to ensure compatibility with the learning algorithm. Each of these stages plays a crucial role in ensuring that the features supplied to the model are both informative and robust. This manual or semi-automated process has traditionally required significant domain expertise and iterative refinement, especially in catalysis, where subtle changes in ligand architecture or electronic properties can have profound effects on reactivity.

Recent advancements in ML, particularly in deep learning and graph-based models such as graph neural networks (GNNs), have shifted part of the feature extraction process from human-driven engineering to model-driven learning. As previously stated, in these frameworks, molecules are often represented as graphs.

This automated descriptor learning holds promise for reducing bias introduced by manual feature selection, uncovering latent chemical relationships, and generalizing across diverse molecular spaces. However, it also comes with increased computational complexity and reduced interpretability, which are important considerations in scientific applications.

Table 2 summarizes key categories of molecular descriptors commonly employed in catalysis-oriented ML applications, along with representative examples and their functional relevance.

5. Model training, validation and evaluation

To develop ML models that are robust and capable of generalizing beyond the training data, it is essential to divide the available dataset into separate subsets for training and validation.[45] The training set is used to fit the model parameters, while the validation set serves to evaluate the model's performance on unseen data, thereby providing an estimate of its predictive power. Model performance is typically quantified using a set of statistical metrics, each suited to different types of predictive tasks. For regression problems, such as predicting catalytic activity or reaction yield, commonly used metrics include the coefficient of determination (R^2), which indicates the proportion of variance in the target variable that is explained by the model. A higher R^2 value suggests a better fit to the data. Another widely used metric is the mean absolute error (MAE), which measures the average magnitude of the differences between predicted and actual values, providing an interpretable indication of prediction accuracy in the same units as the response variable.

In classification tasks, categorizing catalysts as active or inactive, the accuracy score is often reported. This metric reflects the proportion of correct predictions made by the model out of the total number of predictions. However, in cases of class imbalance, additional metrics such as precision,

Table 2 Molecular descriptors in catalysis

Descriptor type	Example	Interpretation/function
Steric	% V_{bur}	Measures the spatial crowding or steric hindrance near a site
Electronic	HOMO–LUMO gap	Indicates reactivity, electrophilicity, or nucleophilicity
Geometric	Bite angle, torsional angle	Describes the 3D conformation and spatial alignment of ligands
Topological	Molecular fingerprints	Encodes structural connectivity without explicit geometry

recall, and F1-score may also be necessary to capture model performance more comprehensively.

To mitigate the risk of overfitting and obtain a more reliable estimate of model generalizability, cross-validation is frequently employed. In this approach, the dataset is partitioned into multiple train-test splits (folds), and the model is trained and evaluated across these partitions in a systematic manner. This allows for a more robust assessment of model stability and ensures that performance metrics are not overly dependent on a particular data split.

In ML, models are validated through a tiered approach: the training set is used to build the model, the test set evaluates its performance on unseen data from the same study, and a separate external validation set, often new experimental data, challenges the model to generalize to completely novel conditions. The high R^2 values reported across all three tiers strongly suggest the model has captured genuine physical insights and is not overfitted, underscoring its predictive reliability.

6. Navigating chemical space: ML strategies for catalytic reaction development

The application of ML in organometallic catalysis has emerged as a powerful tool to accelerate the development and optimization of synthetic reactions. The selection of optimal reaction conditions, including the choice of catalyst, solvent, reagent, temperature, and time, is a fundamental aspect of organic synthesis, critically impacting both the efficiency and selectivity of chemical transformations. Traditionally, this optimization has relied heavily on the intuition and experience of synthetic chemists, often requiring labour-intensive and time-consuming trial-and-error experimentation. Recent advances in ML have introduced transformative methodologies to accelerate and rationalize this process. Several studies have demonstrated that data-driven models can significantly reduce the experimental burden by accurately predicting reaction outcomes under varied conditions, even in the absence of prior empirical data.

Despite significant progress, current methods for predicting reaction conditions still face important limitations. Most approaches cannot reliably predict the full set of reaction parameters, including catalysts, solvents, reagents, and temperature, across large collections of reactions. Moreover, most methods do not account for the interdependence between these factors, which means that compatibility issues between chemicals and reaction conditions are often ignored. Evaluating the performance of these models on large datasets is also challenging, partly because comprehensive, machine-readable databases with standardized classifications of catalysts, solvents, and reagents are lacking, and partly because it is difficult to quantitatively assess predictions that involve complete sets of conditions. Another key challenge lies in representing the chemical context in a way that is both general enough to be widely applicable and specific enough to capture meaningful chemical detail. Representations that are too general may overlook important functional characteristics, while overly specific approaches, such as copying entire reaction conditions from previous experiments, do not provide insights into chemical similarity or allow for broader predictions.

A lot of studies focused narrowly on isolated elements of the chemical context (*e.g.*, solvent choice or catalyst class) or on restricted reaction families. For example, solvent selection has been extensively investigated as an independent problem. Struebing *et al.* combined quantum mechanical (QM) calculations with a computer-aided molecular design framework to identify solvents capable of accelerating reaction kinetics.[46] Although effective in specific cases, this approach is difficult to scale due to the high computational cost associated with QM calculations.

Data-driven methods have also been applied to suggest conditions for particular reaction types. In this context, Marcou *et al.* developed an expert system to predict suitable catalysts and solvents for Michael additions,[71] trained on 198 known reactions.[47] The task was framed as multiple binary classification problems, determining whether a given solvent or catalyst would be appropriate for a particular Michael reaction. Nevertheless, in an external test set, only 8 out of 52 reactions had both predicted solvent and catalyst matching the experimental conditions. Many more studies could be mentioned here, highlighting the common trend of investigating only a few reaction parameters at a time, which often limits general applicability and overlooks the interplay between different components of the chemical context.

6.1 Reaction conditions optimization *via* machine learning models

One of the most direct and practical applications of machine learning (ML) in catalysis is the optimization of reaction conditions. By predicting promising conditions with fewer experimental runs, ML offers an efficient alternative to the traditional trial-and-error approach that often dominates synthetic chemistry.[49]

In 2018, Jensen and colleagues introduced a pioneering study that expanded the scope of ML beyond narrow problems such as solvent or catalyst selection.[50] Instead, their hierarchical neural network simultaneously predicted catalysts, solvents, reagents, and reaction temperature, addressing the multifactorial interplay that governs catalytic outcomes. The authors constructed a hierarchical neural network model trained on approximately 10 million single-step, single-product reactions curated from the Reaxys database. Reactions were encoded using Morgan circular fingerprints for reactants and products, from which reaction fingerprints were derived to represent the structural changes taking place. In other words, the authors trained their model on millions of published reactions and taught it to recognize how molecules change during a transformation, so that the algorithm could then learn which catalysts, solvents, and reagents are typically used under similar circumstances: the model was able to develop its own internal ''map'' of chemical species, where solvents and

reagents with similar roles or properties were placed closer together.

To train the model, the different condition parameters (catalysts, solvents, reagents) were treated as classification problems: for each category, the algorithm had to choose the correct option among hundreds of possible chemicals. Each chemical was represented using a ''one-hot encoding,'' meaning that every candidate was assigned its own unique slot in a long vector, where the correct entry is marked as '1' and all the others as '0'. Put simply: imagine you have a list of 200 possible solvents: one-hot encoding means creating a column for each solvent and marking ''1'' only for the one that was actually used, while all the others get ''0''. In this way, the algorithm knows exactly which solvent is present without assuming any similarity between them. Temperature, on the other hand, is not a yes/no choice but a number that can vary continuously, so it was handled as a regression problem where the model tries to predict a realistic numerical value. A sequential prediction scheme was adopted, in which the catalyst prediction informed the solvent prediction, which in turn conditioned the reagent and temperature predictions. This design mimics the logical reasoning employed by practicing chemists and captures interdependencies among condition elements. Because the dataset was so large, the authors first had to clean and simplify it. Many chemical species appeared only very rarely in the database and including them would have made the model harder to train without adding much useful information. To avoid this ''data sparsity'', species used fewer than 100 times were excluded, leaving a still impressive 11.4 million reactions across 803 catalysts, 232 solvents, and 2247 reagents.

Conditions were modelled as classification problems with one-hot encodings, while temperature was treated as a regression task. A sequential prediction strategy was adopted: catalyst prediction informed solvent choice, which in turn conditioned reagent and temperature outputs, mirroring chemists' decision-making.

Performance was strong: solvents and reagents were correctly ranked among the top ten suggestions *ca.* 83% of the time, catalysts above 90%, and the entire context matched literature conditions in 57% of cases (rising with functionally equivalent alternatives). Temperature predictions were within ± 20 °C in most cases, with accuracy improving when the chemical context was correct. Qualitatively, the network often suggested either the reported conditions or chemically reasonable substitutes, such as predicting piperidine instead of morpholine for Fmoc deprotection. Importantly, inference was in the order of milliseconds per reaction, vastly outperforming nearest-neighbour searches. The study highlighted how ML can both capture functional similarities between species and scale efficiently, offering an invaluable tool for integration into synthesis planning platforms.

With the aim of addressing the dataset challenge, Li and colleagues presented in 2024 AutoTemplate, a data preprocessing protocol designed to enhance the quality and reliability of chemical reaction datasets used in machine learning applications for organic chemistry.[51] Recognizing that the performance of models in yield prediction, retrosynthesis, and reaction condition recommendation depends heavily on dataset integrity, the authors propose a two-stage framework consisting of generic template extraction and template-guided reaction curation. Using simplified SMARTS representations, AutoTemplate derives broadly applicable reaction templates that are then systematically applied to validate, correct, and complete reaction entries—addressing issues such as missing reactants, incorrect atom mappings, and erroneous reactions.

A distinctive aspect of the method lies in its ability to identify and correct false reactions by leveraging reliable entries within the dataset as self-consistent references. Applied across a variety of reaction types, AutoTemplate demonstrates substantial improvements in dataset quality, offering a stronger foundation for developing accurate machine learning models in chemistry. Through this work, Li and colleagues provide a concrete framework for correcting common errors such as missing reactants and atom-mapping inconsistencies, directly reinforcing the point that dataset quality critically affects ML-driven catalysis and synthesis prediction.

Another influential contribution came from the Doyle group and co-workers, who focused on yield prediction in Pd-catalyzed Buchwald–Hartwig aminations.[14] The central challenge was the unpredictable effect of functional groups, such as isoxazoles, that can poison catalysts or trigger side reactions. To tackle this, instead of directly varying substrates bearing heterocycles, they adopted a Glorius fragment additive screening approach that enables the testing of hundreds, if not thousands, of distinct interactions within a short timeframe and with limited resources, effectively transforming a traditionally slow and labour-intensive process into a rapid and streamlined screening strategy. It is a clever screening trick to quickly test if a specific functional group will cause problems in a chemical reaction. Instead of synthesizing complex molecules, scientists simply add a small, representative fragment of the functional group to the reaction mixture. If the reaction proceeds well, the group is compatible. If the yield plummets, the fragment is likely to interfere, perhaps by poisoning the catalyst or initiating a side reaction. It is a high-throughput shortcut to map out what works and what does not. Although not without limitations (free fragments may sometimes behave differently than when incorporated into larger molecular frameworks) it nonetheless provides an exceptional starting point for mapping the landscape of chemical reactivity. This strategy enabled the generation of an unusually rich dataset of over 4600 nanomole-scale reactions, spanning multiple aryl halides, ligands, bases, and additive combinations.

A key methodological advance was the automated generation of 120 quantum-chemical descriptors for all reaction components, allowing the ML model to uncover relevant physicochemical trends without human bias in descriptor choice. Among tested algorithms, Random Forests provided the best performance, achieving $R^2 = 0.92$ on the test set with RMSE = 7.8%. Remarkably, even with only 5% of the dataset, the model outperformed linear regression trained on 70%. The model generalized well, correctly predicting yields for eight unseen

isoxazole additives. Importantly, analysis of descriptor importance revealed that electrophilic isoxazoles inhibit the reaction by undergoing competitive oxidative addition with Pd(0), a mechanistic insight subsequently validated experimentally. This study exemplifies how ML can transform empirical reactivity.

A further advance was made two years later by Fu *et al.*, who developed a deep neural network for the Suzuki–Miyaura cross-coupling reaction, a cornerstone of C–C bond formation.[18] Despite its versatility, this transformation is notoriously difficult to optimize due to the interdependence of catalyst, substrate, and condition choices. The DNN software integrated both molecular descriptors and external variables such as temperature, residence time, and catalyst loading, features often neglected in prior ML studies. Training data comprised 387 well-curated reactions, systematically varying halides, boron reagents, and eight Pd pre-catalysts.[52]

Input features were 44 quantum-mechanical descriptors, including HOMO/LUMO energies, Mulliken charges, bond lengths, and exposed surface areas, providing a physically meaningful basis for prediction. The model achieved excellent performance with $R^2 = 0.945$ on test data and > 0.92 on external validation (see Model Training, Validation, and Evaluation section), substantially outperforming *k*-nearest neighbour regression. Computationally, it was highly efficient, screening over 15 000 candidate reactions in about one minute.

Importantly, predictions were experimentally validated. For one benchmark reaction, the DNN-guided conditions improved yields from 30% to nearly 90% in flask-scale experiments. For new combinations generated by crossovers of known reactants, predicted yields closely matched experimental results (errors often $< 5\%$). Even for structurally modified substrates absent from the training set, the model maintained strong predictive power (see Table 2). Feature analysis underscored the importance of temperature and catalyst loading, while descriptors such as halogen bond length and boron atom charge aligned with established chemical intuition. By linking ML predictions to mechanistic principles, this work bridged the gap between ''black box'' computation and interpretable chemistry.

In 2021, Ebi and co-workers introduced a collaborative ML framework aimed at supporting chemists in the design of reaction conditions.[53] The study was motivated by a central obstacle in applying ML to synthesis: while most models require large and uniform datasets, experimental data are typically scarce, heterogeneous, and often biased toward particular research goals. To overcome this, the authors proposed a system where chemists can contribute data incrementally, with ML algorithms updating predictions iteratively as new results become available. This framework was tested on two widely studied but condition-sensitive transformations: Suzuki–Miyaura cross-couplings and Buchwald–Hartwig aminations.

The core of their method was a Bayesian optimization algorithm that balanced exploration (searching undersampled regions of condition space) with exploitation (refining promising leads). Reaction conditions were encoded by descriptors of ligands, catalysts, bases, and solvents, and the model used acquisition functions to select the next set of experiments predicted to be most informative. By adopting this active-learning approach, the system reduced the experimental effort needed to locate high-yielding conditions compared with traditional grid searches. In practice, the model was able to identify near-optimal conditions in a fraction of the experimental runs that exhaustive screening would have required.

To evaluate its generality, the authors compared the performance of the ML-driven strategy with conventional experimental design in benchmark studies. The Bayesian optimization model consistently converged more rapidly to high-yielding conditions, particularly when multiple parameters interacted non-linearly. Importantly, the method was designed for collaborative use: separate research groups could maintain local autonomy while pooling partial datasets, allowing the algorithm to benefit from a broader data foundation without requiring full centralization of experimental information. This opens a path toward ''distributed intelligence'' in synthesis optimization. The study also highlighted practical limitations: predictions depended strongly on the diversity of the initial dataset. Models trained on narrow or biased data could overfit and mislead exploration.

Across the five studies reviewed, ML has consistently demonstrated its value in optimizing catalytic reactions, albeit through different strategies. Gao *et al.* showed that hierarchical neural networks can successfully predict complete reaction contexts, catalyst, solvents, reagents, and temperature, highlighting ML's ability to learn chemically meaningful representations from large datasets.[50]

Li and Chen developed a two-stage preprocessing framework that extracts and applies simplified SMARTS-based reaction templates to correct dataset errors such as missing reactants and atom-mapping inconsistencies.[51]

The Doyle group demonstrated that random forest models trained on high-throughput experimental data can accurately predict yields in Buchwald–Hartwig aminations and identify inhibitory functional groups, providing both predictive power and mechanistic insight.[14] Fu *et al.* established that deep neural networks integrating quantum mechanical descriptors and reaction parameters can optimize Suzuki–Miyaura cross-couplings, with experimental validation confirming dramatic yield improvements.[18] Ebi *et al.* advanced a collaborative Bayesian optimization framework that reduces the number of required experiments and enables data sharing across laboratories, underscoring the practicality of active learning in data-scarce environments.[53]

In a recent contribution from our group, we reported the application of a ML-driven optimisation strategy to the platinum-catalysed reduction of amides under hydrosilylation conditions.[54] Instead of relying on conventional trial-and-error protocols, we employed the Sunthetics ML platform, an algorithm specifically designed to iteratively refine experimental conditions while minimising the number of required trials. The optimisation targeted the reduction of *N*,*N*-dimethylacetamide to the corresponding amine using 1,1,3,3-tetramethyldisiloxane (TMDS) as reductant and a set of five Pt(II)-based catalysts,

including thioether- and NHC–ligated complexes. The Sunthetics ML platform was employed to iteratively propose reaction conditions by varying three parameters, reaction time (1–24 h), temperature (25–80 °C), and catalyst loading (0.01–1.0 mol%), while continuously updating predictions based on experimentally measured conversion, turnover number (TON), and turnover frequency (TOF). Through only a limited number of experiments, the algorithm identified $[Pt(DMS)_2Cl_2]$ and $[Pt(THT)_2Cl_2]$ as the most promising systems, achieving full conversion under exceptionally mild conditions. Notably, $[Pt(DMS)_2Cl_2]$ at 0.01 mol% loading and 39 °C afforded quantitative reduction within 2 h, corresponding to a TOF of 5002 h^{-1}, one of the highest reported for this transformation. The ability of the algorithm to converge rapidly on optimal catalytic conditions underscores its value not merely as a tool for efficiency but as a driver of discovery, revealing reactivity patterns that would be difficult to predict solely based on empirical screening.

In 2024, Li and colleagues present an innovative approach for the automatic recommendation of reaction conditions within the framework of computer-aided synthesis planning (CASP).[55] The study introduces a two-stage deep learning model that integrates a multi-label classification network with a ranking model to predict suitable reagents, solvents, and reaction temperatures for chemical transformations. A notable feature of this work is the use of hard negative sampling, which generates fictitious reaction conditions that challenge the model to refine its decision boundaries and improve its robustness in distinguishing favorable from unfavorable reaction contexts.

Trained across ten reaction types—Buchwald–Hartwig cross coupling, Chan–Lam coupling, Diels–Alder reaction, Fischer indole synthesis, Friedel–Crafts acylation, Friedel–Crafts alkylation, Grignard reaction, Kumada coupling, Negishi coupling, and reductive amination—the model achieves 73% top-10 accuracy in retrieving at least one correct set of reaction conditions and predicts temperatures within ±20 °C of experimental values in 89% of test cases. Importantly, the model demonstrates the capability to propose viable alternative reaction conditions beyond the confines of the training dataset, underscoring its potential to inspire novel synthetic strategies and accelerate discovery in chemical research.

By enabling the generation and prioritization of diverse reaction conditions based on predicted relevance scores, Li's model represents a significant advancement toward the integration of reaction condition prediction into CASP systems.

Together, these studies (summarized in Table 3) converge on a clear conclusion: ML not only accelerates the search for optimal conditions but also offers interpretable insights into the underlying chemistry, bridging data-driven prediction with

Table 3 Examples of reaction condition optimization *via* machine learning models

Study (Year)	Optimization problem addressed	ML approach	Key results & conclusions
Gao *et al.* (2018)	Recommend complete reaction contexts (catalyst, solvents, reagents, temperature) across diverse organic reactions.	Hierarchical neural networks, Morgan fingerprints; sequential multi-output classification + temperature regression.	Top-10 accuracy ≈83% for solvents/reagents and >90% for catalysts; full-context top-10 ≈57%; temperature MAE ≈25.5 °C (≈19.4 °C when context correct). Millisecond-scale inference; scalable, general recommendations.
Ahneman *et al.* (2018)	Predict yields and diagnose inhibitory functional groups in Pd-catalysed Buchwald–Hartwig aminations using HTE data.	Random Forest regression trained on >4600 nanomole-scale reactions; ∼120 computed descriptors.	Test R^2 ≈ 0.92 (RMSE ≈ 7.8%); strong data efficiency; accurate out-of-sample predictions. Identifies electrophilic isoxazoles as inhibitors *via* competitive oxidative addition (validated).
Fu *et al.* (2020)	Optimize Suzuki–Miyaura cross-coupling conditions incl. catalyst loading, temperature, residence time.	Deep neural network with 44 QM descriptors for reactants and Pd pre-catalysts.	R^2 ≈0.945 (test), >0.92 (external). Screens ∼15 200 candidates in ∼1 min (GPU). Experimental validation: yield ↑ from ∼30% to ∼89%; generalizes to new/cross-over substrates; highlights importance of temperature & catalyst loading.
Ebi *et al.* (2021)	Cut experiments by guiding condition selection for Suzuki–Miyaura and Buchwald–Hartwig with scarce/fragmented data.	Collaborative active learning *via* Bayesian optimization; iterative updates as labs contribute data.	Reaches high-yielding conditions with far fewer runs than exhaustive screening; supports multi-lab pooling. Performance depends on diversity of seed data; narrow/bias can mislead.
Nolan (2024)	Reduction of *N*,*N*-dimethylacetamide with Pt-catalyst; optimization of catalyst identity, loading, temperature, and reaction time.	Sunthetics ML platform; iterative optimisation algorithm adjusting experimental conditions based on conversion, TON, and TOF to minimise the number of experiments.	Identified $[Pt(DMS)_2Cl_2]$ as optimal catalyst; achieved quantitative conversion with 0.01 mol% loading at 39 °C in 2 h, corresponding to TOF = 5002 h^{-1}.
Li (2024)	Prediction and ranking of optimal reaction conditions (reagents, solvents, temperatures) in Buchwald–Hartwig, Chan–Lam, Diels–Alder, Fischer indole synthesis, Friedel–Crafts acylation, Friedel–Crafts alkylation, Grignard reaction, Kumada coupling.	A two-stage deep learning model combining multi-label classification. Hard negative sampling to refine decision boundaries and improve accuracy in challenging reaction contexts.	73% top-10 accuracy for correct reagent/solvent sets and predicts temperatures within ±20 °C in 89% of cases. It successfully suggests multiple viable and novel reaction conditions, demonstrating strong potential for integration into CASP to enhance synthesis planning and reaction optimization.

mechanistic understanding. Together, these studies illustrate complementary strengths of ML-guided optimization. ML models not only reproduce known conditions but also extrapolate to novel substrates, enabling rapid, scalable, and generalizable predictions. Through feature importance analysis, these models often uncover chemically interpretable trends, bridging data-driven insights and mechanistic understanding. A key advantage is efficiency: ML can screen thousands of reaction conditions *in silico*, drastically reducing experimental effort. Active-learning strategies further enhance efficiency by intelligently exploring chemical space, while models trained on high-throughput data accelerate discovery and reveal key reactivity drivers.

Challenges remain, however, including dependence on high-quality, diverse training data; the ''black box'' nature of many algorithms, despite advances in descriptor design; and issues of reproducibility and accessibility due to reliance on specialized or proprietary data. Every approach reported in the literature highlights how ML can transform experimental design: not by replacing the chemist, but by augmenting decision-making, reducing wasted effort, and uncovering opportunities that might otherwise remain hidden.

6.2 Machine learning for predicting enantioselectivity and stereo-control

Predicting how a catalyst will behave toward a given substrate has never been straightforward. Unlike yields, which depend on global thermodynamic or kinetic factors, stereoselectivity arises from subtle differences in transition-state energies, often below 1 kcal mol^{-1}. Capturing such fine energetic balances is notoriously challenging for traditional computational methods and impractical for brute-force experimental screening. A minor modification in the structure of a ligand can transform a highly selective system into a completely inactive one or even generate unexpected catalytic species.

For decades, chemists have tackled these challenges through a combination of intuition, experience, and extensive cycles of trial and error. In recent years, however, machine learning (ML) has emerged as a transformative tool for addressing stereo-control. Strategies range from supervised regression models to unsupervised clustering, each tailored to tackle specific mechanistic challenges.

By learning stereochemical outcomes directly from reaction data, ML models offer the dual advantage of accelerating catalyst discovery and revealing mechanistic principles underlying stereoselectivity. Four recent studies exemplify the multifaceted nature of this approach, highlighting how ML can become a valuable ally for researchers working on stereocontrolled reactions.

A foundational contribution was made by Nandy *et al.* focused on accelerating discovery in transition metal chemistry by applying ML to predict fundamental electronic structure properties, such as HOMO–LUMO of open–shell transition metal complexes.[56] While not directly predicting enantioselectivity, this frontier molecular orbital energetics are crucial for understanding chemical reactivity and dictating optical and electronic properties, which underpin catalytic behaviour. To overcome the challenge of robust and automated data set generation, they introduced the molSimplify automatic design (mAD) workflow and developed topological revised autocorrelation (RAC) descriptors tailored for inorganic chemistry. Their artificial neural network (ANN) models achieved mean absolute errors of 0.15 eV for HOMO level and 0.25 eV for HOMO–LUMO gap, enabling the rapid prediction of properties for diverse complexes and the discovery of molecules with target HOMO–LUMO gaps from a large 15 000-molecule design space in minutes rather than days required by full DFT evaluation. This capability significantly accelerates the initial screening phase, allowing for more efficient exploration of chemical space where subtle structural changes might lead to desired electronic properties. The problem of efficiently navigating complex, multidimensional reaction spaces, traditionally reliant on expert intuition and trial-and-error, has been powerfully tackled by Bayesian optimization (BO).

In 2021, Doyle and colleagues developed a modular framework, experimental design *via* Bayesian optimization (EDBO), and accompanying open-source software to streamline reaction optimization.[57] Recognizing that chemical reaction optimization involves numerous discrete and continuous parameters (*e.g.*, substrate, catalyst, ligand, solvent, temperature, concentration), their BO approach employs a probabilistic surrogate model (Gaussian process) and acquisition functions to intelligently balance the exploration of uncertain areas with the exploitation of promising regions. This strategy ensures the selection of high-quality experimental configurations in fewer evaluations, significantly outperforming human decision-making in both optimization efficiency and consistency. Although their specific applications focused on optimizing reaction yields for palladium-catalysed direct arylation, Mitsunobu, and deoxy-fluorination reactions, the general framework is directly applicable to optimizing enantioselectivity. Their use of DFT-encoded descriptors for reaction components further enhances the model's ability to capture the subtle chemical properties influencing reaction outcomes, thus providing a systematic way to identify optimal stereo-control conditions more rapidly.

Complementing these predictive and optimization tools, Singh and Hernández-Lobato highlighted, in a 2024 contribution, the critical role of understanding biases in existing literature data for successful ML applications in stereocontrol.[58] Their data-driven analysis of transition metal-catalysed asymmetric hydrogenation of olefins (AHO), involving Ir, Rh, and Co-based catalysts, explicitly addressed the challenge that ML models trained on biased data might merely reflect literature trends rather than offer genuine mechanistic insights (Fig. 7). By classifying and visualizing the chemical space of olefins and ligands (using UMAP plots from 2D structural fingerprints), they revealed significant biases in frequently used olefin–ligand combinations and reaction conditions (solvents, temperatures, pressures) for achieving high enantioselectivity. For example, they showed that Ir-catalysed AHO often favours unfunctionalized olefins with P,N-type

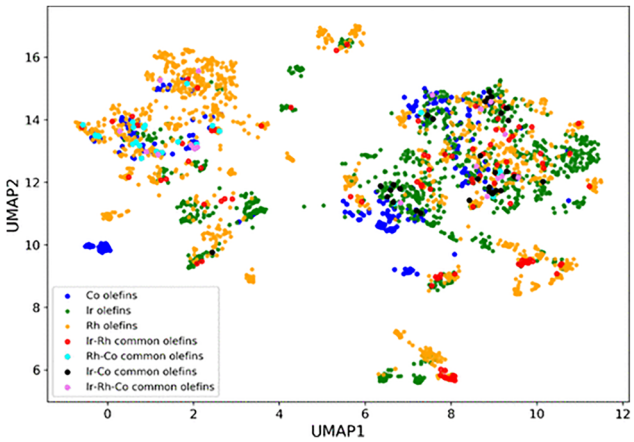

Fig. 7 UMAP plot of the chemical space of olefins used in Ir-, Rh-, and Co-catalyzed asymmetric hydrogenation. The *x*- and *y*-axes correspond to the two UMAP components obtained after dimensionality reduction. Reproduced from ref. 58 with permission from the *J. Org. Chem.*, copyright 2024.

ligands, while Rh catalysis is more suited for functionalized olefins with phosphorus ligands.

Crucially, their work underscored the sparsity of explored olefin–ligand combinations, pinpointing "empty boxes" (Fig. 8) that represent specific olefin–ligand pairs not yet reported or utilized in the extensive literature dataset on asymmetric hydrogenation, thereby revealing unexplored regions of this chemical space.

Notably, these uncharted areas still exhibit considerable potential in terms of median enantioselectivity and therefore warrant further experimental investigation. Such a detailed understanding of data distribution and the underlying structure–selectivity relationships is invaluable for practitioners, as it enables the design of new, diverse, and unbiased reaction datasets, ultimately enhancing both the reliability and predictive power of future ML models aimed at enantioselectivity and stereo-control (Table 4).

The common thread through all these studies is augmenting the chemist's intuition: taken together, these studies illustrate that ML is not a single method but a versatile suite of tools adaptable to various challenges, from predicting fundamental properties and optimizing reactions to uncovering hidden patterns and analysing big data. The common thread is the ability to surpass the limits of human intuition, not by replacing the chemist, but by augmenting their capabilities. By efficiently exploring chemical spaces too vast for traditional approaches, ML acts as a compass: it does not dictate the destination but suggests the most promising directions, reducing the randomness of the discovery pathway and accelerating the development of new stereo-controlled processes.

6.3 Ligand design and screening

The design and screening of ligands are pivotal yet challenging aspects of organometallic catalysis, largely dictating a catalyst's performance, selectivity, and stability. Traditionally, catalyst design has relied heavily on human intuition and local structural searches, often struggling to reconcile multiple conflicting property requirements within the vast chemical space of potential catalysts.

The sheer number of conceivable ligand structures necessitates more efficient property prediction and a deeper understanding of quantitative structure–property relationships. Furthermore, the profound impact of ligand flexibility on molecular properties and catalytic activities has often been overlooked, with systematic quantification of conformational effects remaining underdeveloped. These complexities underscore the need for advanced, data-driven approaches, such as machine learning (ML), to navigate the chemical space effectively and accelerate the discovery of novel and efficient catalysts.

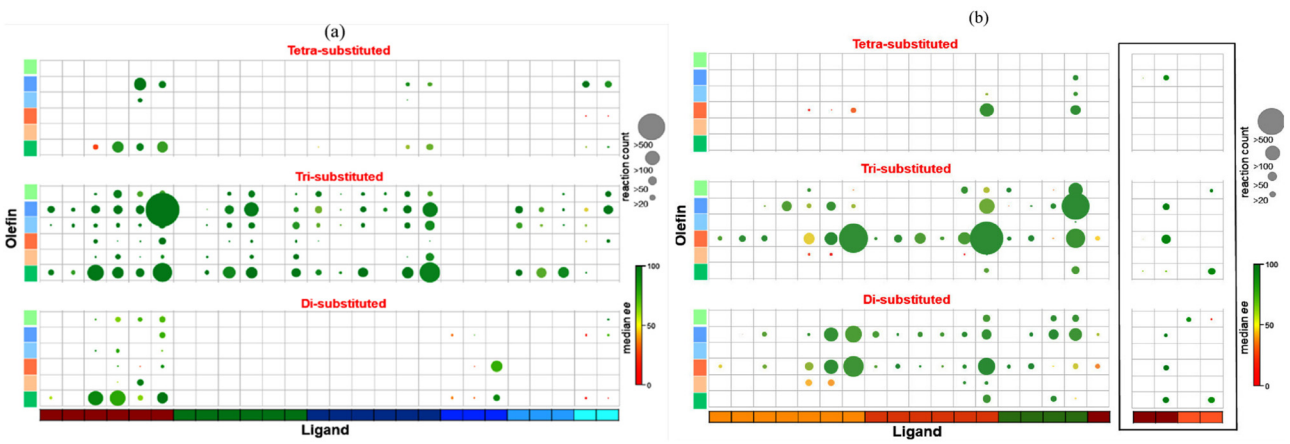

Fig. 8 Plots of olefin–ligand combinations for (a) Ir, (b) Rh, and Co (enclosed in a black box)-catalysed asymmetric hydrogenation. The *y*-axis corresponds to olefin type where each colour displays the identity of olefin. The *x*-axis represents the type of ligand. The circle size corresponds to the number of reactions. The colour corresponds to the median enantioselectivity of all reactions in the given category. Reproduced from ref. 58 with permission from the *J. Org. Chem.*, copyright 2024.

Table 4 Examples of machine learning models for predicting enantioselectivity and stereo-control

Study (Year)	Optimization problem addressed	ML approach	Key results & conclusions
Nandy *et al.* (2018)	Property prediction in transition-metal complexes (spin states, HOMO–LUMO gaps) as a foundation for stereochemical modelling	LASSO regression, Kernel ridge regression, artificial neural networks (ANNs) trained on revised autocorrelation (*RAC*) descriptors	ANNs achieved MAEs as low as 0.15 eV (HOMO), identifying non-local steric descriptors as dominant. Framework demonstrated transferability to stereocontrol by capturing ligand environment effects.
Shields *et al.* (2021)	Optimization of Pd-catalyzed C–H arylation and Buchwald–Hartwig couplings, with potential extension to stereoselective outcomes	Bayesian optimization (Gaussian process surrogate models with DFT descriptors; EDBO platform)	Outperformed expert chemists and DOE in finding optimal conditions with fewer experiments. Balance of exploration/exploitation enabled systematic optimization of selectivity-driven transformations.
Hueffel *et al.* (2021)	Catalyst speciation problem: predicting ligand-induced formation of dinuclear Pd(I) *versus* Pd(0)/Pd(II) species in cross-coupling	Unsupervised ML (*k*-means clustering on ligand knowledge base of 348 phosphines; refinement with DFT-derived problem-specific descriptors)	Identified clusters of ligands favouring Pd(I) dimers, reducing search space from 348 to ~25%. Experimental validation yielded 8 new air-stable Pd(I) dimers (some previously unknown). Showcased ML's ability to reveal non-intuitive ligand–speciation relationships.
Singh & Hernández-Lobato (2024)	Prediction of enantioselectivity in asymmetric hydrogenation of olefins (Ir, Rh, Co catalysts; >12 000 reactions)	Random forests, gradient boosting, neural networks trained on substrate/ligand fingerprints and reaction conditions	Models achieved $R^2 > 0.8$ for % ee in Ir/Rh systems. Identified mechanistic trends: Ir favours minimally functionalized olefins; Rh requires coordinating groups. Highlighted dataset bias toward high % ee as key limitation.

One significant advancement in this area is the kraken platform, developed by Sigman group in 2022, which addresses the challenge of comprehensively mapping the chemical space of organophosphorus ligands and facilitating their design and optimization for catalysis.[59] The platform was created to move beyond empirical approaches by providing a comprehensive discovery platform for monodentate organophosphorus(III) ligands. A key innovation of Kraken is its generation of comprehensive physicochemical descriptors based on representative conformer ensembles, thereby explicitly accounting for ligand conformational flexibility, a feature previously underdeveloped in ligand characterization. Using quantum-mechanical (QM) methods, Sigman and colleagues calculated 190 descriptors for an initial set of 1558 ligands, including commercially available examples and highly cited structures, forming "Virtual Library 1" (VL1).[59] To drastically expand the explored chemical space, they developed ML models, including a "Bag of Substituents" (BoS) model and more generalizable approaches using molecular fingerprints and graph convolutional neural networks (Fig. 10).

These models were trained on the QM data to predict properties for over 300 000 new ligands (VL2, based on unary and binary substituent combinations) and enabled on-demand queries for approximately 191 million entries (VL3, covering ternary combinations). The platform employs dimensionality reduction techniques like Uniform Manifold Approximation and Projection (UMAP) and Principal Component Analysis (PCA) to visualize the property space, aiding in the identification of unexplored regions and the understanding of property limits.[59]

Notably, Kraken demonstrated its utility in inverse catalyst design by building linear free energy relationships and regression models from experimental data, which could then predict the performance of the entire ligand database to suggest optimal candidates for subsequent experiments. This capability was exemplified by its application to enantiospecific Pd-catalyzed sp^3–sp^2 cross-coupling reactions, where it identified structurally unique ligands optimal for the reactions and even suggested hybrid designs.

Building on the utility of ML in navigating ligand space, Schoenebeck and colleagues recently tackled the specific challenge of rationally designing multinuclear catalysts, where the correlation between a metal's ligand and its preferred speciation (oxidation state, geometry, and nuclearity) is often unknown.[60] Their work focused on accelerating the identification of suitable ligands to form trialkyl phosphine-derived dihalogen-bridged Ni(I) dimers, a class of complexes previously unexplored for bulky trialkyl phosphine ligands, and difficult to access through conventional methods from common precursors like $Ni(COD)_2$.

To overcome the absence of known Ni(I) dimer references, Karl *et al.* employed an assumption-based unsupervised machine learning approach. They started with a subspace of 66 ligands from an existing database (LKB-P) known to be similar to bulky trialkyl phosphines, like $P(t\text{-Bu})_3$. They then introduced "pseudo-positive" ($P(t\text{-Bu})_3$) and "pseudo-negative" (tri(neopentyl)phosphine) references to guide the clustering algorithm, based on the assumption that these would (or would not) lead to the desired Ni(I) dimer geometry.[60]

Problem-specific descriptors, 184 in total, encompassing steric and electronic effects of free ligands and various Ni complexes, derived from DFT calculations, were then generated and refined. Through sequential *k*-means clustering, the algorithm successfully grouped 16 phosphine ligands, many of which had not been previously utilized in Ni catalysis, alongside $P(t\text{-Bu})_3$ as promising candidates for Ni(I) dimer

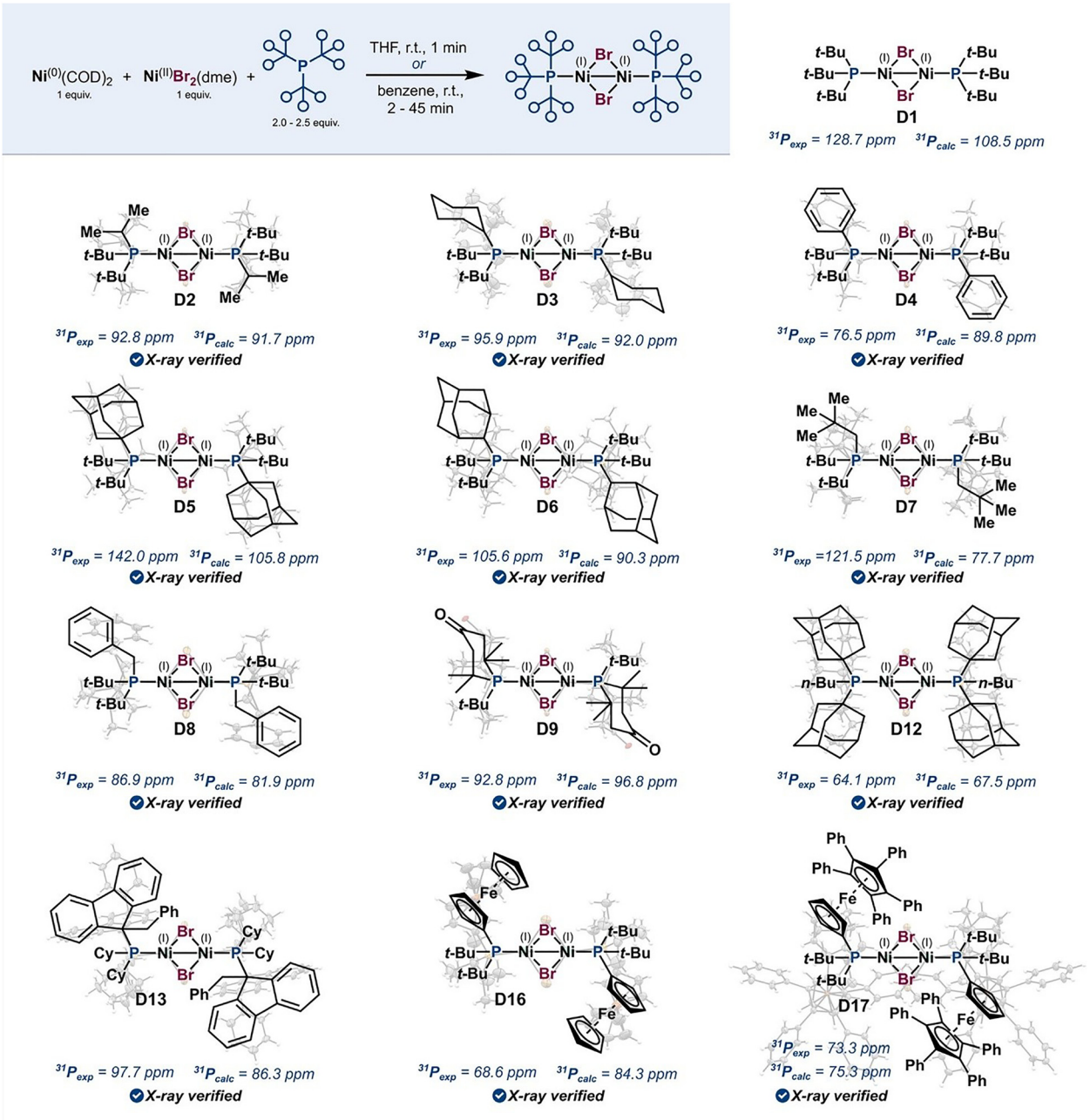

Fig. 9 Structures of the 12 novel dibromo-bridged Ni(I) dimers (D1–D13, D16, and D17) confirmed by X-ray crystallography. These complexes were identified using an unsupervised ML workflow combining *k*-means clustering of a phosphine ligand database with DFT-derived descriptors, guiding the discovery of previously unexplored Ni(I) dimers. Selected dimers were validated experimentally *via* comproportionation, highlighting the ML-driven exploration of new Ni(I) chemistry. Reproduced from ref. 54 with permission from the *J. Am. Chem. Soc.*, copyright 2023.

formation.[60] Experimental validation confirmed the ML predictions, leading to the successful synthesis and X-ray crystallographic characterization of 12 novel dibromo-bridged Ni(I) dimers (see Fig. 9). This ML-guided discovery enabled a significant catalytic application: the iodo-selective arylation of polyhalogenated arenes with competing C–Br and C–Cl sites, achieved in under 5 minutes at room temperature with low catalyst loading, a feat previously unmet by alternative mono- or dinuclear Pd or Ni catalysts. This study powerfully demonstrates the utility of ML in navigating unexplored ligand spaces for specific metal speciation. Further refining our understanding of ligand performance, Doyle and Sigman addressed the challenge of ''reactivity cliffs'' in catalysis, where a binary outcome (reaction occurs or not) depends on a critical threshold value of a molecular descriptor (Fig. 11).[61] They aimed to identify a physically meaningful descriptor that could classify catalyst ligation states and predict reaction outcomes.

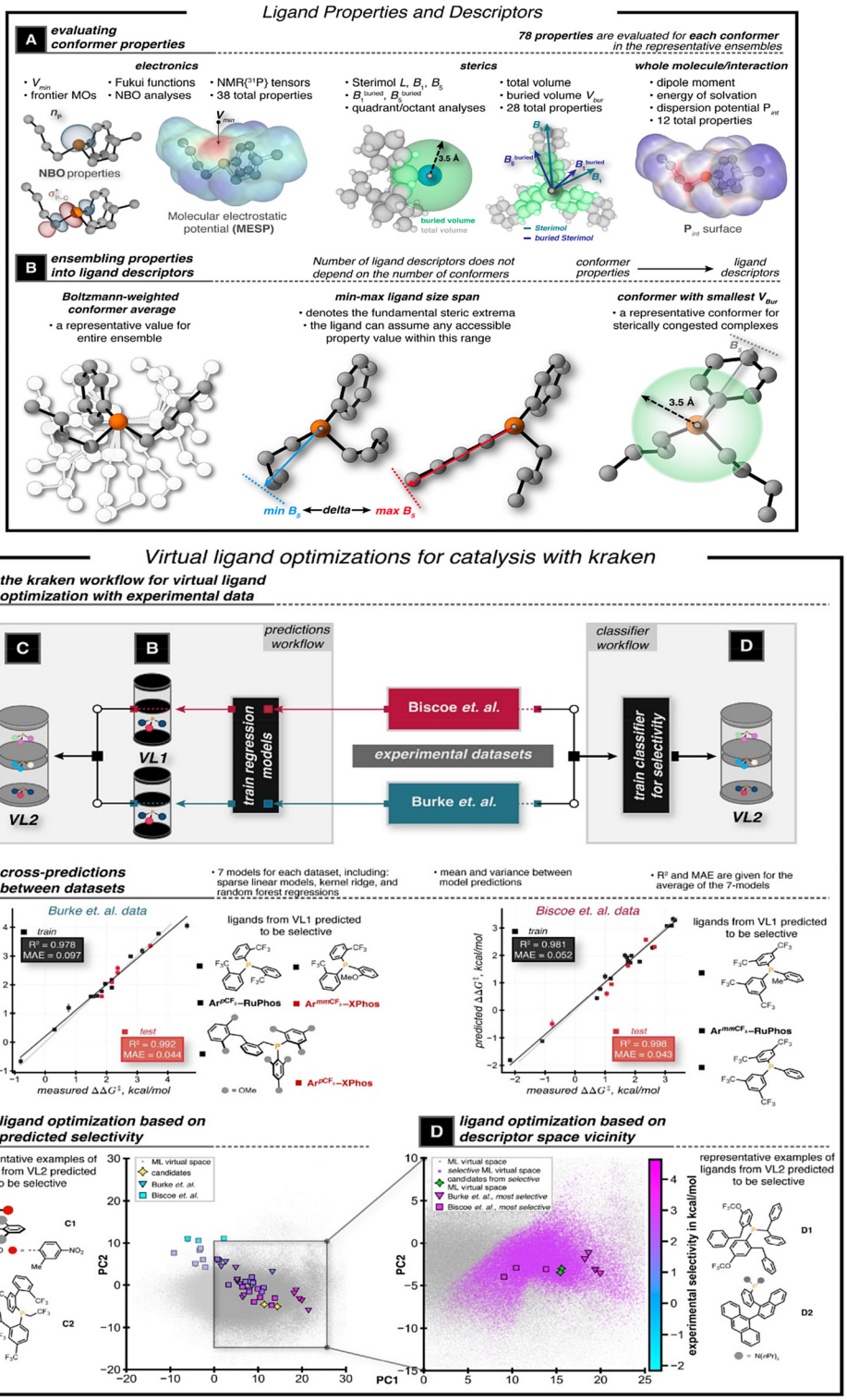

Fig. 10 (A) Above: illustrations of some properties computed for each conformer. (B) Ensembling conformer properties to generate ligand descriptors. Note that absolute buried volume in Å^3 is used in this library instead of the more common percent buried volume % V_{bur} (% V_{bur} = V_{bur}/1.8) to retain comparability with the total volume. Below: using kraken for virtual ligand optimizations in asymmetric catalysis. (A) General workflow for the case study. (B) Statistical modelling of experimental results to predict how data from one reported reaction could inform ligand choice in the other through a virtual screen of VL1 for ligands that are predicted to result in high selectivity for the stereo-retentive cross-coupling. (C) Combining the statistical models for both reactions to evaluate the entirety of VL2 for new selective ligands. (D) Exploring the PCA descriptor space to determine ligands with novel structures in the high-selectivity regime. Reproduced from ref. 59 with permission from JACS, copyright 2022.

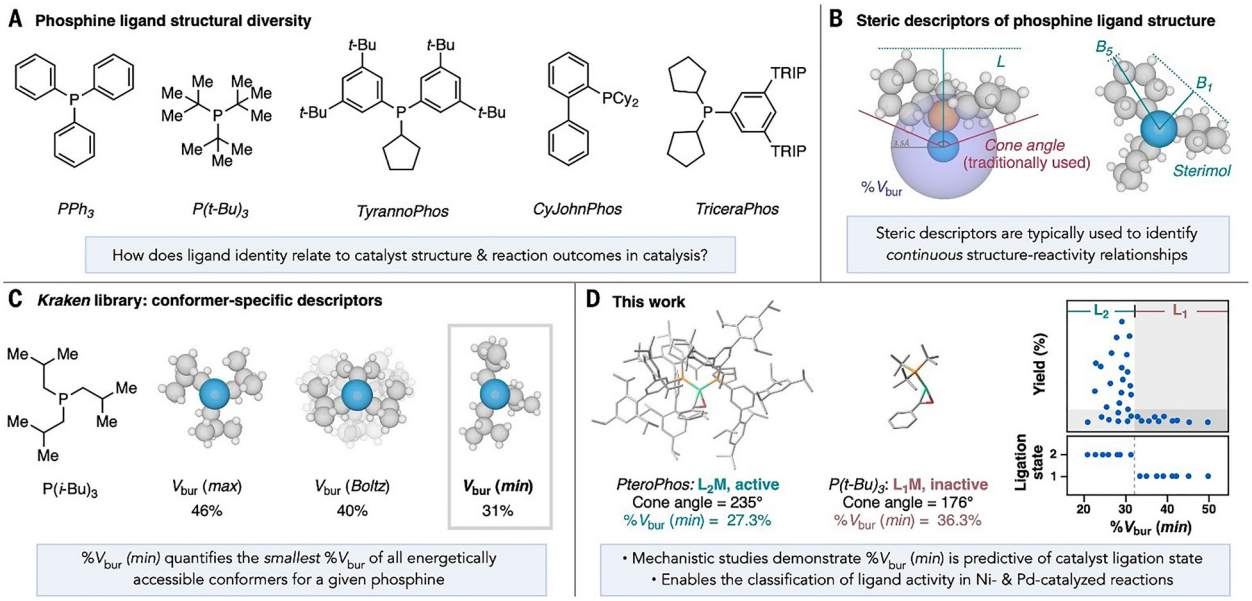

Fig. 11 (A) Examples of monodentate phosphines used in Ni and Pd cross-coupling reactions, including TyrannoPhos and TriceraPhos (DinoPhos ligands) recently reported by the Doyle lab. TRIP, 2,4,6-triisopropylphenyl. (B) Commonly used methods for quantifying phosphine steric properties. Cone angle is the traditionally used descriptor for quantifying monodentate phosphine steric bulk and is defined as the angular width (in degrees) of an imaginary cone needed to encapsulate the entire phosphine structure; the vertex of the cone is defined by a metal atom bound to the ligand with a bond length of 2.28 Å. % V_{bur}, a more modern descriptor designed initially to study N-heterocyclic carbenes, is defined as the volume percent of the phosphine's atoms that fill an imaginary sphere of 3.5 Å radius that is centered on a metal atom bound to the phosphine with a bond length of 2.28 Å. Sterimol descriptors B1 and B5 describe the lowest and highest width of the ligand perpendicular to the metal–phosphorus axis, respectively, and Sterimol L describes the ligand's length along that axis. (C) Phosphine descriptor library (kraken) capturing multiple ligand conformers, with maximum, Boltzmann average, and minimum % V_{bur} values of the conformational ensemble of P(i-Bu)$_3$ shown. (D) This work. L1, one equivalent of ligand bound to metal; L2, two equivalents of ligand bound to metal; M, metal. Reproduced from ref. 61 with permission from Science, copyright 2021.

While traditional descriptors like the Tolman cone angle often fail to capture the nuanced topological features and conformational flexibility that influence reactivity, Doyle and Sigman utilized the comprehensive descriptor set from the kraken platform.[61] Their key finding was the identification of minimum percent buried volume (% V_{bur} (min)) as a distinctive steric descriptor capable of classifying reactivity cliffs in 11 Ni- and Pd-catalyzed cross-coupling datasets. This descriptor quantifies the smallest % V_{bur} among all energetically accessible conformers of a ligand, effectively representing the steric bulk within the metal's first coordination sphere under optimal fitting conditions.

Through spectroscopic and crystallographic organometallic studies, they demonstrated that a % V_{bur} (min) threshold of approximately 32% accurately predicted the binary outcome of bis-ligated (L2) *versus* monoligated (L1) metal complexes. This was mechanistically validated by DFT calculations showing a sharp decrease in the free energy of ligand dissociation (DGdissoc) for [L2Ni(benzaldehyde)] complexes as % V_{bur} (min) approached 32%, corresponding to a significant increase in Ni–P bond length due to steric pressure. The universality of this concept was shown by applying the classification workflow to various Pd-catalyzed cross-coupling reactions, where % V_{bur} (min) thresholds were observed, reflecting the distinct L1 or L2 requirements of different reactions. This work provided a robust, quantitative tool for mechanistically rationalizing ligand performance, predicting catalyst ligation state, and even identifying scenarios where steric properties are not rate-determining, pointing towards ligandless reactivity or nanoparticle formation.[61] These studies collectively showcase the transformative power of ML in organometallic catalysis for ligand design and screening. From the comprehensive mapping and expansion of ligand chemical space by the kraken platform, through the targeted discovery of ligands for specific metal speciation using unsupervised ML, to the identification of critical reactivity cliffs and mechanistic insights offered by % V_{bur} (min), ML-driven workflows are proving instrumental in overcoming traditional limitations. These approaches not only accelerate the discovery of novel and efficient ligands but also deepen our fundamental understanding of structure–reactivity relationships, paving the way for more informed and rational catalyst design.

6.4 Mechanism prediction and pathway elucidation

Understanding reaction mechanisms remains a foundational challenge in catalysis, as it is crucial for optimizing catalytic processes and designing more efficient catalysts. The detailed knowledge of how reactants transform into products allows for better control over reaction outcomes, helping to increase reaction rates, selectivity, and sustainability. While density functional theory (DFT) has traditionally served this purpose by providing accurate models of molecular interactions, ML offers complementary capabilities. ML can efficiently navigate complex or multistep catalytic cycles, identify patterns and

predict reaction pathways that might be too intricate for traditional methods. By integrating ML with DFT, researchers can accelerate the discovery of new catalytic processes and improve the design of catalytic systems.

In this context, Roet and colleagues introduced in 2021 a novel ML-based method designed to enhance our understanding of how chemical reactions occur in molecular simulations, with a particular focus on liquid systems.[62] The technique leverages decision tree (DT) classifiers to pinpoint key features-mainly atomic distances-that play a critical role in facilitating a chemical reaction. Unlike traditional methods that rely on 3D atomic positions, which can be influenced by rotations or translations of molecules, this approach reformulates simulation data into a distance matrix format. This adaptation enhances the robustness of the analysis and makes it more compatible with rare-event simulation methods like replica exchange transition interface sampling (RETIS), which tracks rare reaction events without the need for predefined reaction coordinates. To further refine the results, the authors employ random forests and statistical averaging to estimate uncertainties in their predictions. Additionally, the method can be extended to incorporate other descriptors such as velocities or angles, allowing for more flexibility in future applications.

In a practical demonstration of the method, the authors studied proton transfer in formic acid (FA) surrounded by four or six water molecules, representing a simple but crucial reaction in acid–base chemistry (Fig. 12).

Their results revealed that the number of water molecules significantly influences both the rate and diversity of the reaction pathways. When only four water molecules were present, the reaction was highly specific, requiring precise alignment of FA and specific atomic distances below defined thresholds. If these conditions were met, there was a 71% chance that the system would follow a reactive pathway. In contrast, with six water molecules, the system exhibited much greater flexibility, with multiple combinations of atomic distances enabling proton transfer. Though the reactivity remained similar (72%), the reaction happened approximately 10 million times faster compared to the case with fewer water molecules. This shift in reaction dynamics highlights how the number of surrounding water molecules influences the overall structure and hydrogen bonding network, which in turn affects the speed and variety of reaction pathways. The method proved to be computationally efficient, training quickly and scaling well with the amount of data, but its accuracy was highly dependent on the quality of the input data. If biased or low-quality simulations were used, the ML model could highlight misleading features. Nonetheless, the approach offers a clear, interpretable, and transferable tool for investigating complex reaction mechanisms.

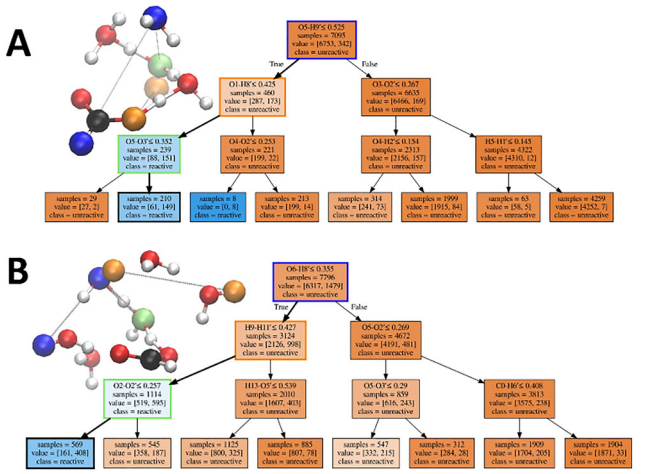

Fig. 12 Decision trees for the systems with (A) four water molecules or (B) six water molecules around the formic acid. Reproduced from ref. 62 with permission from the American Chemical Society, copyright 2023.

In a separate effort, Baldi and colleagues presented a new ML framework for predicting the detailed mechanisms of organic chemical reactions.[63] Traditionally, such predictions have relied on rule-based systems or expert chemists' intuition. However, the authors proposed a data-driven approach that focuses on analysing the interactions between molecular orbitals (MOs), specifically the interactions between electron donors and acceptors in reactant molecules. This framework considers various reaction conditions, such as temperature and solvent type, and addresses several challenges inherent in modelling chemical reactions. First, there was a lack of a suitable dataset for training the model. To overcome this, the authors used a rule-based system called Reaction Explorer, which simulates reactions and identifies productive steps (those leading to final products). By simulating over 6 million reactions, they identified 2989 productive reactions and labelled them according to whether the atoms involved participated in a productive step. This dataset, along with labels for "reactive" and "non-reactive" atoms, served as the foundation for training their ML model.

The second challenge the team faced was the sheer number of possible reactions, which grows rapidly due to the many potential combinations of molecular orbital pairs. To handle this complexity, the authors trained two separate neural networks to classify atoms in reactants as likely to be reactive under certain conditions. These networks used over 1500 chemical and structural features, such as atomic charge, neighbouring atoms, and bond arrangements, to assess the reactivity of individual atoms. The classification step effectively reduced the number of reactions considered by 94%, while maintaining a very low false negative rate of under 0.1%. This ensured that the model could accurately classify reactions without missing productive ones. The third challenge was ranking the remaining reactions by their likelihood of success. The authors treated this as a ranking problem, using pairwise comparisons to score reactions based on their productivity. A neural network trained for this task helped identify the most likely reaction mechanisms, achieving impressive accuracy. In nearly 90% of cases, the model ranked the correct mechanism first, and in 99.9% of cases, all productive reactions were found within the top five predictions.

The authors demonstrated the model's ability to predict detailed reaction mechanisms with several examples, such as

the Claisen condensation reaction, a classic organic reaction in which an ester undergoes an intramolecular transformation to form a new carbon–carbon bond. The model also showed its ability to generalize beyond the training data, correctly predicting reactions that involved larger molecules, such as seven-membered ring formations. In another example, the model compared two competing reaction mechanisms involving an oxonium intermediate and ranked them according to their likelihood of success. Although the model slightly misranked one mechanism as more favourable, this was chemically justifiable, as hydrogen transfer reactions are often reversible and may be preferred under certain conditions. This highlights the model's potential for providing detailed mechanistic insights in organic chemistry.

In addition to predicting reaction outcomes, the system was also capable of predicting the type of reaction (polar, radical, or pericyclic), allowing for more nuanced predictions (Fig. 13).[64] Following the classification and ranking steps, the model was able to generate multi-step synthetic routes using a depth-first search algorithm, successfully finding plausible reaction pathways, including sequences involving protecting group formations and classic synthetic reactions like the Robinson annulation (Fig. 13(B)). The authors argue that their system offers a powerful tool for both research and education in chemistry. The system is publicly available through a web-based interface, making it accessible to the broader scientific community.

In 2023, Schaaf and De introduced an active learning protocol for developing ML force fields (MLFFs) to model catalytic reactions at the atomic scale.[65] Their method aims to overcome the limitations of traditional density functional theory (DFT) methods, which are accurate but computationally expensive. By combining DFT data with ML, the protocol efficiently predicts minimum energy paths (MEPs) for catalytic reactions. It was applied to the hydrogenation of CO_2 to methanol on indium oxide, and the results demonstrated the model's ability to reproduce reaction intermediates and transition states with high accuracy (Fig. 13(C)). The approach also allowed for the identification of previously unrecognized rate-limiting steps and provided more realistic free energy profiles by incorporating thermal and entropic effects. This framework could significantly reduce computational costs while maintaining high accuracy, opening new possibilities for detailed mechanistic studies in catalysis and materials science.

In another application of ML, Sui and Zhao developed two models for optimizing peracetic acid (PAA)-based advanced oxidation processes (AOPs) for environmental water treatment.[66] These processes are important for degrading recalcitrant organic pollutants, but traditional methods can be expensive and produce harmful by-products.[67] The CRCO-ML model (CRCO = catalyst and reaction condition optimization), trained on a dataset of over 1000 experimental cases, predicted the most influential factors affecting PAA activation, such as catalyst composition, dosage, and environmental conditions (Fig. 14). The model was validated experimentally, and its predictions showed high accuracy, with errors below 10%. The MI-ML model (MI = mechanism identification) focused on identifying the degradation mechanisms of pollutants, using quantum chemical descriptors and quenching experiments. This model successfully identified key reactive oxygen species, such as hydroxyl radicals and organic radicals, which play a critical role in the degradation process. The study demonstrates the potential of ML to optimize environmental processes, although the authors stress the need for larger datasets and more interpretable models.

Finally, in 2024, Reiher and colleagues introduced STEERING WHEEL, a novel and flexible algorithm designed to guide the automated exploration of chemical reaction networks (CRNs).[68] CRNs-graph-based representations of chemical transformations involving nodes for compounds and reactions-are crucial for understanding complex catalytic mechanisms. Traditionally, building such networks through first-principles calculations has been both time-consuming and computationally demanding, often limited by the lack of universal, scalable methods. Fully automated tools can be efficient but tend to oversimplify, while semi-automated approaches require expert intervention and do not scale well. The STEERING WHEEL addresses these limitations by combining the autonomy of algorithmic exploration with intuitive user control, offering a dynamic balance between automation and expert guidance.

Embedded in the SCINE software suite and operated through its user-friendly interface HERON, the STEERING WHEEL enables real-time interaction with the evolving chemical space. Users can alternate between network expansion steps-which introduce new reactions and intermediates-and selection steps, which filter, prioritize, or redirect the search based on criteria such as structural motifs, reactive site types, or computational cost. This hybrid strategy allows chemists to strategically target relevant regions of the reaction space without specifying individual intermediates, while also ensuring reproducibility by requiring each expansion to complete before advancing.

The framework was tested across diverse and increasingly complex systems. In the study of Wilkinson's catalyst, the method successfully reconstructed both the Halpern and Brown mechanisms for olefin hydrogenation,[69] preserving the full triphenylphosphine ligand environment to maintain electronic and steric accuracy (Fig. 15(A)). This level of detail revealed multiple novel isomeric intermediates and agostic interactions not previously documented. While some configurations deviated from expected geometries, density functional theory (DFT) optimizations confirmed the validity of key five-coordinate structures. The exploration also clarified which sterically hindered isomers were thermodynamically unfeasible, supporting their absence in automated outputs.

In a separate application to Ziegler–Natta-catalysed propylene polymerization,[64] the method modelled two polymerization cycles and a termination step, simulating monomer insertions and chain termination *via* β-hydride elimination (Fig. 15(B)).[48] This enabled the identification of expected products like 2-methylpropene and 2,4-dimethylpentene, along with 18 additional hydrocarbon by-products. Although the use

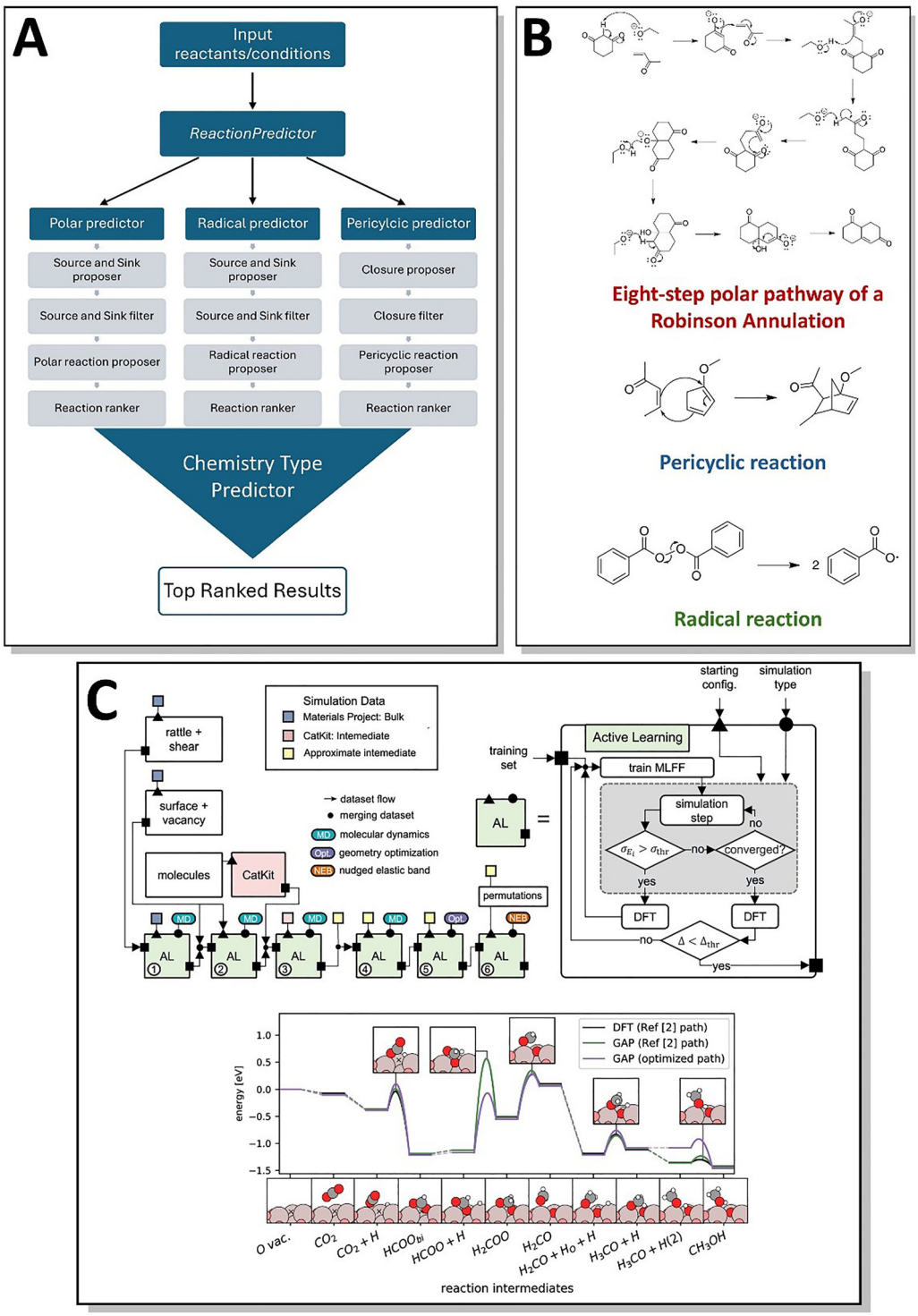

Fig. 13 (A) Workflow diagram and (B) examples of mechanisms interpreted by the ReactionPredictor algorithm. (C) Workflow diagram and reaction profile of the hydrogenation of CO_2 to methanol studied by Schaaf and De. Fig. 2(C) is reproduced from ref. 65 with permission from Nature, copyright 2023 for both.

of the semi-empirical GFN2-xTB model limited energy precision, the protocol demonstrated its capacity to handle growing conformational complexity through the integration of conformer generation tools and post-processing with RDKit and XYZ2MOL.

The approach also proved effective for analysing the Monsanto process,[70] an industrial reaction involving methanol carbonylation catalysed by a rhodium complex.[65] This reaction poses significant challenges due to the presence of multiple intertwined catalytic cycles and the complexity of transition

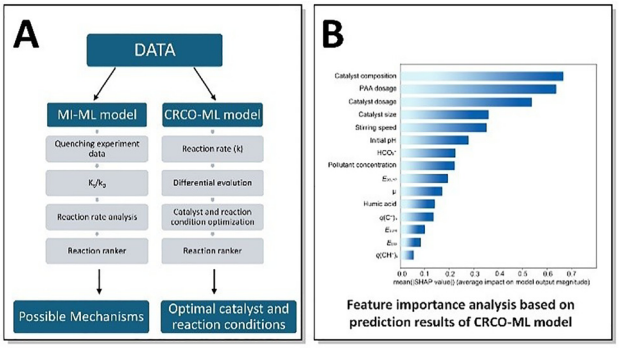

Fig. 14 (A) Architecture of CRCO-ML and MI-ML models. (B) Feature importance from the CRCO-ML model's predictions analysis. Fig. 3(B) is reproduced from ref. 64 with permission from Elsevier, copyright 2023 for both.

metal and solution-phase chemistry (Fig. 15(C)).[71] Using the guided exploration protocol, the authors successfully reconstructed the known mechanism and uncovered new catalytic pathways, including one initially appearing stoichiometric but shown to be catalytic upon additional exploration. The ability to adaptively refine exploration protocols allowed for the detection of mechanistic features previously missed by automated methods, such as the methyl iodide activation step. Here, switching from GFN2-xTB to DFT resolved missing pathways and improved energy profiles, underscoring the importance of high-level methods for accuracy.

In the most complex application, the authors investigated olefin polymerization catalysed by a gallium single-site complex on silica.[72] This system required deeper exploration-spanning 19 network expansion steps-due to the disordered silica surface and the variety of possible hydrocarbon rearrangements. The method reproduced known reaction intermediates, identified unexpected products such as *trans*-butene and 1,3-butadiene, and corrected previously misunderstood steps, including enantioselective transformations. Although more computationally intensive, the guided approach yielded an expansive network of nearly 1800 species and over 14 000 reactions, including degradation and side reaction paths, demonstrating its scalability and effectiveness.

The authors emphasize that while the current methodology allows for exhaustive and adaptive exploration, its modular architecture also supports further integration of advanced tools. Potential extensions include multi-level electronic structure models, automated solvation corrections, expanded conformer generation, and the incorporation of ML, kinetic simulations, and path-based heuristics to automate selection decisions. The infrastructure is fully open-source and freely available through SCINE HERON, promoting transparency, reproducibility, and community-driven development.

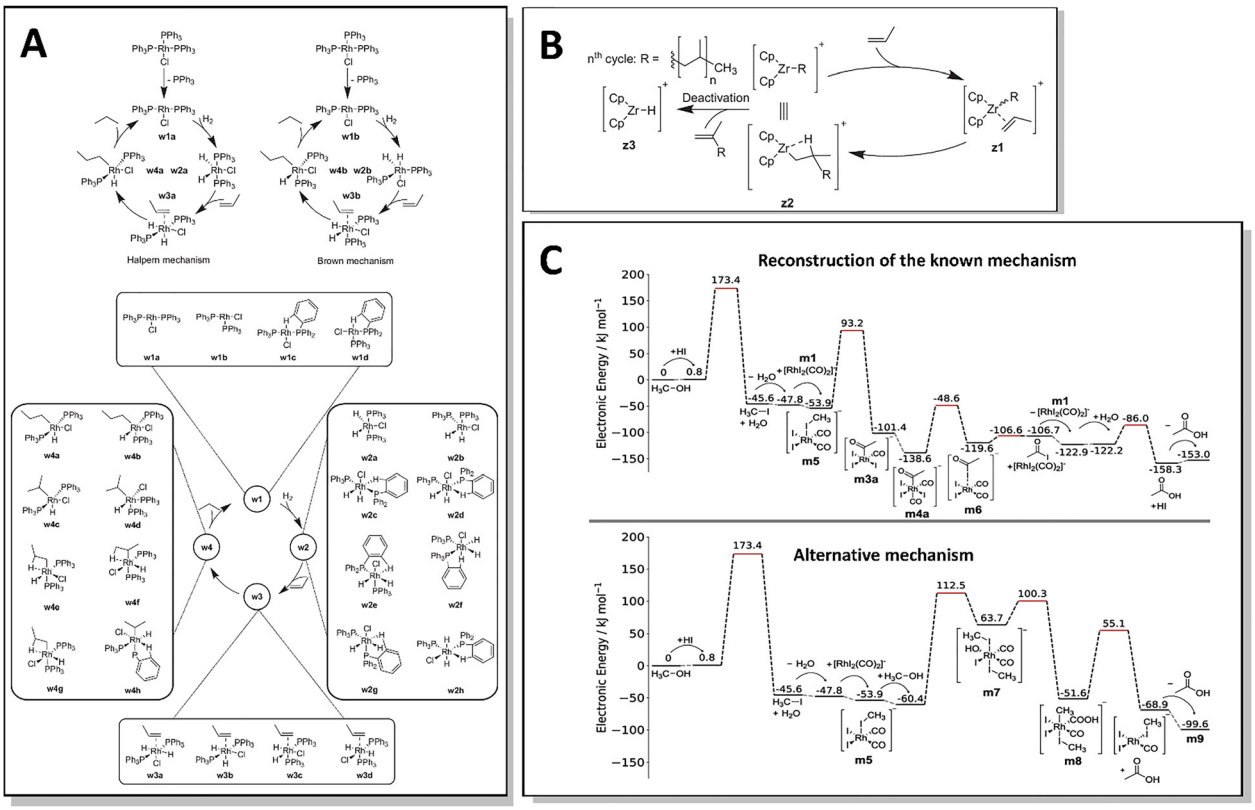

Fig. 15 Examples of mechanisms ((A) Rh-based olefin hydrogenation, (B) Ziegler–Natta propylene polymerization and (C) monsanto methanol carbonylation) interpreted by the STEERING WHEEL algorithm. Reaction schemes are reproduced from ref. 68 with permission from Nature, copyright 2023 for both.

In the broad field of reaction mechanisms, Li and colleagues have systematically evaluated three advanced strategies---transfer learning, delta learning, and feature engineering—to enhance activation energy prediction using graph neural networks (GNNs) trained on low-cost semiempirical quantum mechanical (SQM) data.[73] Using the Chemprop/D-MPNN framework, the authors investigated how each approach balances accuracy and computational cost when high-level data are scarce. Among the tested methods, delta learning proved most effective, accurately mapping low-level SQM activation energies to high-level CCSD(T)-F12a targets while requiring only 20–30% of the high-level data used by other methods. Although delta learning demands computationally intensive transition state searches, it offers remarkable efficiency gains in data-limited contexts. Transfer learning showed variable results depending on the alignment of pretraining and target datasets, whereas feature engineering provided modest improvements, particularly for thermodynamic descriptors. Overall, the study offers practical guidelines for selecting data augmentation strategies in ML-driven reaction engineering and underscores the trade-offs between accuracy, data availability, and computational efficiency in predicting activation energies.

6.5 Discovery of novel catalysts and complexes

The discovery of novel catalysts and complexes is a critical yet challenging endeavor in chemistry, often limited by the vastness of chemical space and the complexity of predicting molecular properties through traditional methods or intuition. ML is a powerful tool to overcome these limitations, significantly accelerating the identification of new functional materials. By training models on computational or experimental data, ML enables the rapid prediction of key properties. This accelerated prediction allows for the systematic exploration and enumeration of vast candidate catalyst spaces, even those encompassing previously unsynthesized ligands or counterintuitive combinations, to uncover unexpected design rules and exceptions to conventional chemical wisdom. Ultimately, ML-driven discovery empowers researchers to identify and validate entirely new complexes with desired properties, expanding the frontiers of catalyst design beyond human-biased approaches.

In homogeneous catalysis, a major challenge is to predict and understand the speciation of metal catalysts, that is, the specific forms they adopt in solution, defined by their nuclearity (*e.g.*, monomer *vs.* dimer), oxidation state, and ligation state. These features are crucial because they directly influence a catalyst's reactivity, efficiency, and selectivity. In the case of palladium (Pd) catalysis, a long-standing question has been why certain ligands promote the formation of dinuclear Pd(I) complexes rather than the more common Pd(0) or Pd(II) species.[10] This knowledge gap has posed a major obstacle to the rational design of such highly effective catalysts. While traditional ligand maps have provided useful insights, they fall short of fully capturing the intricate relationships between ligand characteristics and catalyst speciation. Compounding the problem, machine learning (ML) approaches typically require large experimental datasets, which are rarely available for such complex speciation problems. To overcome these limitations, Schoenebeck and colleagues[10] developed an innovative unsupervised ML workflow that remarkably required only five experimental data points for its successful implementation.[10] Their strategy involved a multi-step approach: initially, a large database of 348 phosphine ligands (LKB-P) was subjected to

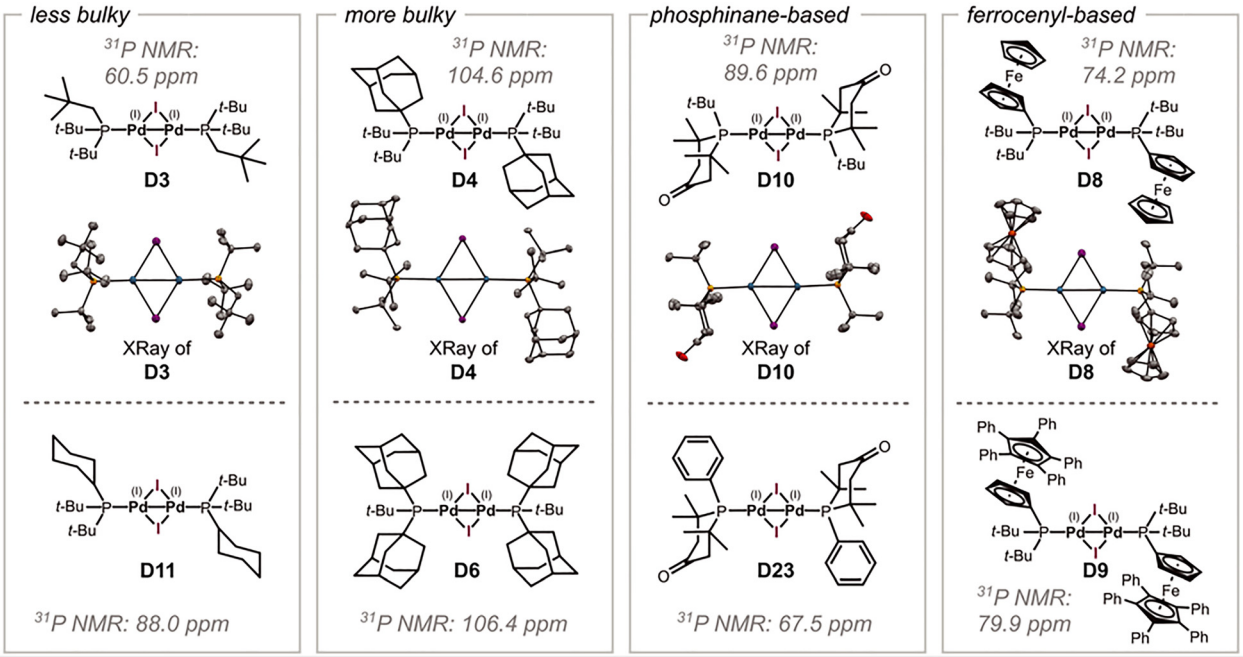

Fig. 16 Newly synthesized Pd(I) dimers and their X-ray crystallographic structures. The figure is from ref. 10 with permission from *Science*, copyright 2021.

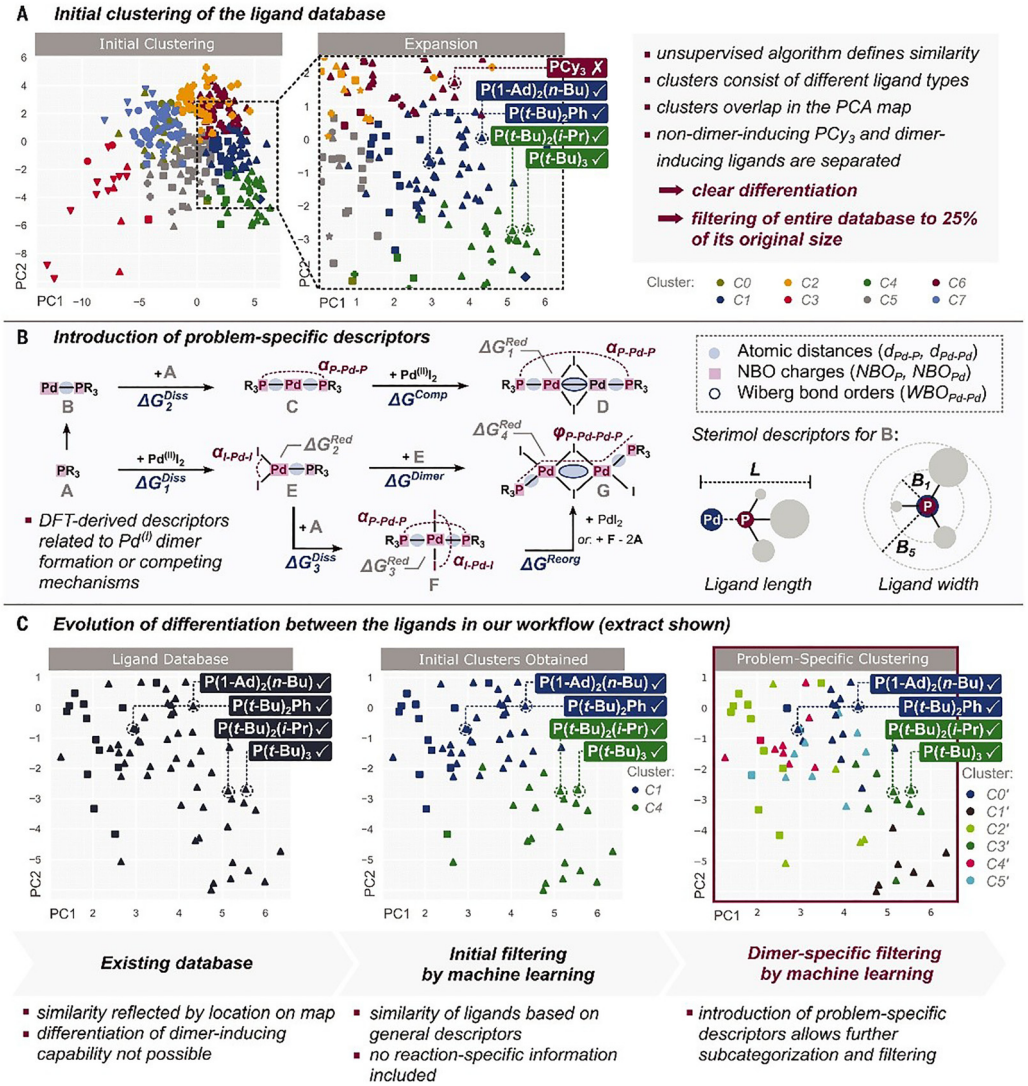

Fig. 17 (A) Initial clustering of the LKB-P using the unsupervised *k*-means algorithm (k = 8; see Fig. S8 and S9 for further visualization). (B) Newly introduced descriptors relating to Pd(I)-dimer formation (see Table S2 for details). (C) Illustration of the differentiation of ligands after initial clustering (middle) and second, problem-specific refinement (right; see Fig. S16 for detailed plot) *versus* original database (left). The same subset of ligands is illustrated. The three schemes are from ref. 10 with permission from Science, copyright 2021.

clustering based on general properties to effectively reduce the vast ligand space (Fig. 17). Crucially, they then introduced 42 problem-specific descriptors, derived entirely *in silico via* DFT calculations. These specialized descriptors were designed to capture the effects of ligands on the stability and geometry of Pd(I) dimer formation, focusing on aspects like Pd–ligand bond properties, conformational effects, electronic charges, and various reaction energies. A second round of clustering, utilizing these problem-specific descriptors, allowed the algorithm to identify 21 promising ligand candidates, including some that had not been previously synthesized. This ML-driven discovery led to the experimental verification and synthesis of eight new air-stable Pd(I) dimers, notably including a previously unknown phosphinane ligand (Fig. 16).

Another significant challenge in catalysis is designing highly selective catalysts for reactions, like in the Doyle work dealing with alkane partial oxidation.[74]

This complexity largely stems from the spin-state-dependent reactivity of metal-oxo intermediates, which makes it difficult to establish robust structure–property relationships using conventional methods. While high-throughput computational screening is powerful, it becomes combinatorially prohibitive when considering multiple metals, spin states, and a wide array of ligands. Furthermore, widely successful approaches in computational screening, such as linear scaling relationships, often prove limited and break down for isolated, under-coordinated metal sites or for spin-state-dependent metal-oxo formation.

Doyle and collaborators addressed these challenges by training Machine Learning (ML) models, specifically Kernel Ridge Regression (KRR) and Artificial Neural Networks (ANNs), to predict spin-state-dependent metal-oxo formation energies (ΔE_{oxo}). For their models, they developed "revised autocorrelations (RACs)" as novel connectivity-only features specifically tailored for inorganic chemistry. KRR models were initially

employed for feature selection and analysis, revealing the dominance of nonlocal, electronic ligand properties in influencing ΔE_{oxo}, a finding that contrasts with previous observations for other transition metal complex properties. Subsequently, ANNs were utilized to enumerate a vast theoretical catalyst space, encompassing over 37 000 candidates. This extensive exploration not only uncovered expected design rules, such as the destabilization of metal-oxo species with increasing d-filling, but also revealed unexpected trends and exceptions, including the orthogonal tunability of oxidative stability and oxo formation energies. Finally, by integrating the ANNs with a genetic algorithm (GA) optimization, the researchers systematically explored this expanded chemical space. This approach led to the discovery of novel and often counterintuitive combinations of metals and oxidation states with unexpected oxo formation energies for oxidatively stable complexes.

These two studies collectively demonstrate the transformative potential of ML in overcoming fundamental challenges in catalyst discovery. From deciphering complex speciation behaviors to navigating vast chemical spaces for novel reactive intermediates, ML provides unprecedented capabilities to identify, predict, and synthesize new catalysts and complexes. By moving beyond intuition and traditional screening, ML accelerates the research pipeline, offering a data-driven path to more efficient, selective, and sustainable catalytic processes. Future research continues to integrate these sophisticated ML methodologies with experimental and computational techniques, paving the way for autonomous catalyst discovery and the elucidation of complex catalytic phenomena.

7. Challenges and future directions

Over the past decade, ML has evolved from a niche tool to a central component of research in organometallic catalysis. From optimizing conditions and predicting yields to uncovering new mechanisms and designing novel ligands, ML has demonstrated its versatility and power. However, despite the field's momentum, several challenges remain that must be addressed to fully realize ML's potential in catalysis. As Sigman and Doyle argue, many datasets used in ML are biased or incomplete, leading to models that lack generalizability.[75] Advances in automated experimentation and standardized data reporting will be crucial to overcome this limitation. Data sparsity, especially for negative or low-yielding reactions, limits model generalization, and many ML models remain ''black boxes'', hindering interpretability and trust.

Furthermore, transferring models across different reaction classes or catalyst families often leads to performance degradation. Fortunately, solutions are beginning to emerge. Active learning frameworks enable models to iteratively query new data, thereby improving themselves over time. Transfer learning and domain adaptation techniques show promise in helping models generalize across diverse chemical spaces. Additionally, interpretability methods, such as SHAP and attention mechanisms, can reveal which features are driving predictions. Looking ahead, the integration of ML with autonomous labs, robotic synthesis, and real-time feedback loops may define the next generation of catalyst discovery and optimization. As datasets continue to expand and models become more sophisticated, ML is poised not only to assist chemists but to transform how catalysis research is conducted. The rapid progress of the field, driven by interdisciplinary collaboration between chemists, data scientists, and engineers, suggests that the next decade will see even greater integration of ML into the workflows of catalytic science, heralding a new era of accelerated, data-driven discovery and innovation.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data included and leading to conclusions presented in this manuscript are included in the manuscript.

Acknowledgements

The authors are grateful to the BOF (starting and senior grants to SPN) as well as the FWO for financial support.

References

1 P. M. Murray, S. N. Tyler and J. D. Moseley, Beyond the Numbers: Charting Chemical Reaction Space, *Org. Process Res. Dev.*, 2013, **17**, 40–46.
2 I. H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, *SN Comput. Sci.*, 2021, **2**, 160.
3 D. Frey, J. H. Shin, C. Musco and M. A. Modestino, Chemically-Informed Data-Driven Optimization (ChIDDO): Leveraging Physical Models and Bayesian Learning to Accelerate Chemical Research, *React. Chem. Eng.*, 2022, **7**, 855–865.
4 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, Graph Neural Networks for Materials Science and Chemistry, *Commun. Mater.*, 2022, **3**, 93.
5 H. Sheikh, C. Prins and E. Schrijvers, Artificial Intelligence: Definition and Background, *Mission AI: The new system technology*, Springer, 2023, pp. 15–41.
6 Y. LeCun, Y. Bengio and G. Hinton, Deep Learning, *Nature*, 2015, **521**, 436–444.
7 N. Artrith, K. T. Butler, F. X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, Best Practices in Machine Learning for Chemistry, *Nat. Chem.*, 2021, **13**, 505–508.
8 J. C. Mitchell and R. Harper, *The Essence of ML*, 1988, pp. 28–46.
9 M. W. Berry, A. Mohamed and B. W. Yap, *Supervised and Unsupervised Learning for Data Science*, Springer, 2020.
10 J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen and F. Schoenebeck, Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning, *Science*, 2021, **374**, 1134–1140.
11 A. K. Jain, J. Mao and K. M. Mohiuddin, Artificial Neural Networks: A Tutorial, *Computer*, 1996, **29**, 31–44.
12 D. Liu, Z. Xu, X. Lu, H. Yu and Y. Fu, Linear Regression Model for Predicting Allyl Alcohol C–O Bond Activity under Palladium Catalysis, *ACS Catal.*, 2022, **12**, 13921–13929.
13 G. Biau and E. Scornet, A Random Forest Guided Tour, *Test*, 2016, **25**, 197–227.

14 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning, *Science*, 2018, **360**, 186–190.
15 J. Dong and S. Hu, The Progress and Prospects of Neural Network Research, *Inf. Control*, 1997, **26**, 360–368.
16 P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen and T. Bligaard, Machine Learning for Computational Heterogeneous Catalysis, *ChemCatChem*, 2019, **11**, 3581–3601.
17 G. Deshmukh, P. Ghanekar and J. Greeley, Deep Learning for Computational Heterogeneous Catalysis: Fundamentals and Applications, *J. Indian Inst. Sci.*, 2025, 1–25.
18 Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan, F. Zhong, D. Wang, X. Luo, K. Chen, H. Liu, J. Wang, H. Jiang and M. Zheng, Optimizing Chemical Reaction Conditions Using Deep Learning: A Case Study for the Suzuki–Miyaura Cross-Coupling Reaction, *Org. Chem. Front.*, 2020, **7**, 2269–2277.
19 D. E. Rumelhart and J. L. McClelland, PDP Research Group, *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*, The MIT press, 1986.
20 D. Marr, Approaches to Biological Information Processing: Physics and Mathematics of the Nervous System. Proceedings of a Summer School, Trieste, Italy, Aug. 1973. M. Conrad, W. Güttinger, and M. Dal Cin, Eds. Springer-Verlag, New York, 1974. Xiv, 584 p., Illus. Paper, $18.50. Lecture Notes in Mathematics, vol. 4. Science, 1975, 190, 875–876.
21 A. Krenker, J. Bešter and A. Kos, Introduction to the Artificial Neural Networks, *Artif. Neural Netw. Methodol. Adv. Biomed. Appl. InTech*, 2011, 1–18.
22 F. Rosenblatt, The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychol. Rev.*, 1958, **65**, 386.
23 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning Internal Representations by Error Propagation*, 1985.
24 L. Personnaz, I. Guyon and G. Dreyfus, Collective Computational Properties of Neural Networks: New Learning Mechanisms, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1986, **34**, 4217.
25 S. I. Gallant, Perceptron-Based Learning Algorithms, *IEEE Trans. Neural Netw.*, 1990, **1**, 179–191.
26 P. D. Wasserman, *Neural Computing: Theory and Practice*, Van Nostrand Reinhold Co., 1989.
27 I. Aleksander and H. Morton, *An Introduction to Neural Computing*, Van Nostrand Reinhold Co., 1990.
28 S. E. Fahlman, *An Empirical Study of Learning Speed in Back-Propagation Networks*, Carnegie Mellon University, Computer Science Department Pittsburgh, PA, USA, 1988.
29 A. Corma, J. M. Serra, E. Argente, V. Botti and S. Valero, Application of Artificial Neural Networks to Combinatorial Catalysis: Modeling and Predicting ODHE Catalysts, *ChemPhysChem*, 2002, **3**, 939–945.
30 K. Omata and M. Yamada, Prediction of Effective Additives to a Ni/Active Carbon Catalyst for Vapor-Phase Carbonylation of Methanol by an Artificial Neural Network, *Ind. Eng. Chem. Res.*, 2004, **43**, 6622–6625.
31 Z.-Y. Hou, Q. Dai, X.-Q. Wu and G.-T. Chen, Artificial Neural Network Aided Design of Catalyst for Propane Ammoxidation, *Appl. Catal., A*, 1997, **161**, 183–190.
32 T. R. Cundari, J. Deng and Y. Zhao, Design of a Propane Ammoxidation Catalyst Using Artificial Neural Networks and Genetic Algorithms, *Ind. Eng. Chem. Res.*, 2001, **40**, 5475–5480.
33 T. Umegaki, Y. Watanabe, N. Nukui, K. Omata and M. Yamada, Optimization of Catalyst for Methanol Synthesis by a Combinatorial Approach Using a Parallel Activity Test and Genetic Algorithm Assisted by a Neural Network, *Energy Fuels*, 2003, **17**, 850–856.
34 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K. R. Müller, Schnet–a Deep Learning Architecture for Molecules and Materials, *J. Chem. Phys.*, 2018, **148**, 241722.
35 J. Gasteiger, J. Groß and S. Günnemann, Directional Message Passing for Molecular Graphs, *arXiv*, 2020, preprint, arXiv:200303123, DOI: 10.48550/arXiv.200303123.
36 Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu and A. Comprehensive, Survey on Graph Neural Networks, *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, **32**, 4–24.
37 Z. Wang, W. Li, S. Wang and X. Wang, The Future of Catalysis: Applying Graph Neural Networks for Intelligent Catalyst Design, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2025, **15**, e70010.
38 R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, A Survey of Methods for Explaining Black Box Models, *ACM Comput. Surv. CSUR*, 2018, **51**, 1–42.
39 Y. Nohara, K. Matsumoto, H. Soejima and N. Nakashima, Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital, *Comput. Methods Programs Biomed.*, 2022, **214**, 106584.
40 S. M. Lundberg and S. I. Lee, A Unified Approach to Interpreting Model Predictions, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4765–4774.
41 X. Wang, Y. Jin, S. Schmitt and M. Olhofer, Recent Advances in Bayesian Optimization, *ACM Comput. Surv.*, 2023, **55**, 1–36.
42 E. Brochu, V. M. Cora and N. A. De Freitas, Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, *arXiv*, 2010, preprint, arXiv:10122599, DOI: 10.48550/arXiv.10122599.
43 A. Jalali, J. Azimi and X. Fern, Exploration *vs.* Exploitation in Bayesian Optimization, *arXiv*, 2012, preprint, arXiv:1204.0047, DOI: 10.48550/arXiv.1204.0047.
44 P. Feliot, J. Bect, E. Vazquez and A. Bayesian, Approach to Constrained Single-and Multi-Objective Optimization, *J. Glob. Optim.*, 2017, **67**, 97–133.
45 H. Clavier and S. P. Nolan, Percent Buried Volume for Phosphine and N-Heterocyclic Carbene Ligands: Steric Properties in Organometallic Chemistry, *Chem. Commun.*, 2010, **46**, 841–861.
46 T. Burzykowski, M. Geubbelmans, A. J. Rousseau and D. Valkenborg, Validation of Machine Learning Algorithms, *Am. J. Orthod. Dentofac. Orthop.*, 2023, **164**, 295–297.
47 H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Siougkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, Computer-Aided Molecular Design of Solvents for Accelerated Reaction Kinetics, *Nat. Chem.*, 2013, **5**, 952–957.
48 G. Marcou, J. Aires de Sousa, D. A. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, Expert System for Predicting Reaction Conditions: The Michael Reaction Case, *J. Chem. Inf. Model.*, 2015, **55**, 239–250.
49 L.-Y. Chen and Y.-P. Li, Machine learning-guided strategies for reaction conditions design and optimization, *Beilstein J. Org. Chem.*, 2024, **20**, 2476–2492.
50 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, Using Machine Learning To Predict Suitable Conditions for Organic Reactions, *ACS Cent. Sci.*, 2018, **4**, 1465–1476.
51 L.-Y. Chen and Y.-P. Li, AutoTemplate: enhancing chemical reaction datasets for machine learning applications in organic chemistry, *J. Cheminf.*, 2024, **16**, 74.
52 M. A. Düfert, K. L. Billingsley and S. L. Buchwald, Suzuki-Miyaura Cross-Coupling of Unprotected, Nitrogen-Rich Heterocycles: Substrate Scope and Mechanistic Investigation, *J. Am. Chem. Soc.*, 2013, **135**, 12877–12885.
53 T. Ebi, A. Sen, R. N. Dhital, Y. M. A. Yamada and H. Kaneko, Design of Experimental Conditions with Machine Learning for Collaborative Organic Synthesis Reactions Using Transition-Metal Catalysts, *ACS Omega*, 2021, **6**, 27578–27586.
54 E. Casillo, B. P. Maliszewski, C. A. Urbina-Blanco, T. Scattolin, C. S. J. Cazin and S. P. Nolan, Machine Learning Directed Discovery and Optimisation of a Platinum-Catalysed Amide Reduction, *Chem. Commun.*, 2024, **60**, 14597–14600.
55 L.-Y. Chen and Y.-P. Li, Enhancing chemical synthesis: a two-stage deep neural network for predicting feasible reaction conditions, *J. Cheminf.*, 2024, **16**, 11.
56 C. Nandy, C. Duan, J. P. Janet, S. Gugler and H. J. Kulik, Strategies and software for machine learning accelerated discovery in transition metal chemistry, *Ind. Eng. Chem. Res.*, 2018, **57**, 13973–13986.
57 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. Martinez Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis, *Nature*, 2021, **590**, 89–96.
58 S. Singh and J. M. Hernández-Lobato, Data-Driven Insights into the Transition-Metal-Catalyzed Asymmetric Hydrogenation of Olefins, *J. Org. Chem.*, 2024, **89**, 12467–12478.
59 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 and A. Aspuru-Guzik, A Comprehensive Discovery Platform for

Organophosphorus Ligands for Catalysis, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
60 T. M. Karl, S. Bouayad-Gervais, J. A. Hueffel, T. Sperger, S. Wellig, S. J. Kaldas, U. Dabranskaya, J. S. Ward, K. Rissanen, G. J. Tizzard and F. Schoenebeck, Machine Learning-Guided Development of Trialkylphosphine $Ni^{(I)}$ Dimers and Applications in Site-Selective Catalysis, *J. Am. Chem. Soc.*, 2023, **145**, 15414–15424.
61 S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman and A. G. Doyle, Univariate Classification of Phosphine Ligation State and Reactivity in Cross-Coupling Catalysis, *Science*, 2021, **374**, 301–308.
62 S. Roet, C. D. Daub and E. Riccardi, Chemistrees: Data-Driven Identification of Reaction Pathways via Machine Learning, *J. Chem. Theory Comput.*, 2021, **17**, 6193–6202.
63 M. A. Kayala and P. Baldi, A Machine Learning Approach to Predict Chemical Reactions, in *Advances in Neural Information Processing Systems*, ed. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, Curran Associates, Inc., 2011, vol. 24.
64 M. A. Kayala and P. Baldi, ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning, *J. Chem. Inf. Model.*, 2012, **52**, 2526–2540.
65 L. L. Schaaf, E. Fako, S. De, A. Schäfer and G. Csányi, Accurate Energy Barriers for Catalytic Reaction Pathways: An Automatic Training Protocol for Machine Learning Force Fields, *npj Comput. Mater.*, 2023, **9**, 180.
66 W. Zhuang, X. Zhao, Q. Luo, X. Lv, Z. Zhang, L. Zhang and M. Sui, Task Decomposition Strategy Based on Machine Learning for Boosting Performance and Identifying Mechanisms in Heterogeneous Activation of Peracetic Acid Process, *Water Res.*, 2024, **267**, 122521.
67 C. Ochs, K. Garrison, P. Saxena, K. Romme and A. Sarkar, Contamination of Aquatic Ecosystems by Persistent Organic Pollutants (POPs) Originating from Landfills in Canada and the United States: A Rapid Scoping Review, *Sci. Total Environ.*, 2024, **924**, 171490.
68 M. Steiner and M. Reiher, A Human-Machine Interface for Automatic Exploration of Chemical Reaction Networks, *Nat. Commun.*, 2024, **15**, 3680.
69 J. C. Vantourout, L. Li, E. Bendito-Moll, S. Chabbra, K. Arrington, B. E. Bode, A. Isidro-Llobet, J. A. Kowalski, M. G. Nilson and K. M. Wheelhouse, Mechanistic Insight Enables Practical, Scalable, Room Temperature Chan–Lam *N*-Arylation of *N*-Aryl Sulfonamides, *ACS Catal.*, 2018, **8**, 9560–9566.
70 P. Cossee, Ziegler-Natta Catalysis I. Mechanism of Polymerization of α-Olefins with Ziegler-Natta Catalysts, *J. Catal.*, 1964, **3**, 80–88.
71 D. Forster, On the Mechanism of a Rhodium-Complex-Catalyzed Carbonylation of Methanol to Acetic Acid, *J. Am. Chem. Soc.*, 1976, **98**, 846–848.
72 Y. Xu, N. J. LiBretto, G. Zhang, J. T. Miller and J. Greeley, First-Principles Analysis of Ethylene Oligomerization on Single-Site Ga^{3+} Catalysts Supported on Amorphous Silica, *ACS Catal.*, 2022, **12**, 5416–5424.
73 H.-C. Chang, M.-H. Tsai and Y.-P. Li, Enhancing Activation Energy Predictions under Data Constraints Using Graph Neural Networks, *J. Chem. Inf. Model.*, 2025, **65**, 1367–1377.
74 A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman and H. J. Kulik, Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation, *ACS Catal.*, 2019, **9**, 8243–8255.
75 P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, Dataset Design for Building Models of Chemical Reactivity, *ACS Cent. Sci.*, 2023, **9**, 2196–2204.

