

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

## Machine Learning in Homogeneous Catalysis: First Steps and Best Practices

Journal:	ACS Catalysis
Manuscript ID	cs-2025-064393
Manuscript Type:	Viewpoint
Date Submitted by the Author:	10-Sep-2025
Complete List of Authors:	Dalmau, David; Instituto de Sintesis Quimica y Catalisis Homogenea, Inorganic chemistry García-Abellán, Susana; Instituto de Sintesis Quimica y Catalisis Homogenea, Inorganic Chemistry Alegre-Requena, Juan Vicente; Instituto de Sintesis Quimica y Catalisis Homogenea, Inorganic Chemistry

SCHOLARONE™  
Manuscripts

# Machine Learning in Homogeneous Catalysis: First Steps and Best Practices

David Dalmau,<sup>‡</sup> Susana García-Abellán,<sup>‡</sup> Juan V. Alegre-Requena<sup>\*</sup>

Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC-Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009 Zaragoza, Spain.

**KEYWORDS.** Machine learning, Homogeneous catalysis, Descriptor generation, Data-driven catalyst discovery

## INTRODUCTION

Recent advances in machine learning (ML) offer new opportunities for homogeneous catalysis, from accelerating discovery to optimizing performance and enabling sustainable design.<sup>1,2</sup> However, the development of new catalysts in this domain remains predominantly empirical, with limited integration of data-driven methodologies. As a result, progress is still largely guided by intuition and trial-and-error approaches that differ little from those used decades ago.

One of the main barriers preventing the widespread adoption of ML is educational, as most chemists lack formal training in data science and ML techniques can seem out of their reach.<sup>3,4</sup> In addition, the supporting ecosystem, including curated datasets, user-friendly tools, and clear benchmarking practices, is still underdeveloped or scattered across domains.<sup>5</sup> This situation limits the use of ML to a small number of specialized groups, while its integration into routine experimental workflows is still uncommon.<sup>6</sup>

This Viewpoint offers a set of guidelines and considerations for implementing ML tools in homogeneous catalysis in a way that is both rigorous and accessible. The manuscript is particularly suited for experimental and computational chemists with little to no experience in ML who intend to apply this technology in their research. The protocols discussed are particularly useful for projects involving dozens to hundreds of experiments or calculations, a common scenario in the field.

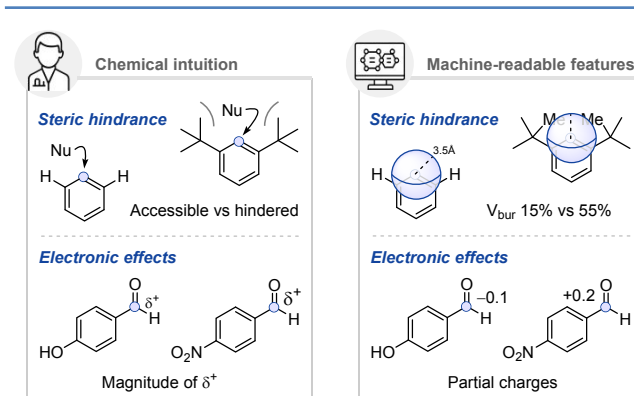
Herein, we outline how to represent molecules in a way that is computer readable, structure datasets, and apply ML models that balance predictive power with interpretability. Through selected case studies, we highlight applications commonly encountered in homogeneous catalysis where chemists can benefit from data-driven strategies, particularly in substrate and catalyst sampling or discovery.

Our aim is to help chemists engage more critically and effectively with ML, not just as tool users, but as informed practitioners aware of its possibilities and limitations. Ultimately, we argue that the thoughtful integration of ML can transform catalysis, not by replacing chemical intuition, but by enhancing it with data-driven insight. While this work focuses on homogeneous catalysis, it is simply an illustration of how coupling ML with chemistry can drive significant breakthroughs across the broader chemistry and materials domains.

## FIRST STEP: DIGITALIZATION OF MOLECULES

One of the first steps in enabling ML predictions is to represent molecules in a way that is computer readable. When chemists examine a molecule, they intuitively recognize functional groups, electronic effects, and steric environments. For example, a nitro group (NO<sub>2</sub>) is electron-withdrawing, a tert-butyl group (tBu) introduces steric hindrance, and a metal center modulates reactivity through its ligand environment. These concepts, while intuitive to humans, are meaningless to an algorithm unless translated into numerical values.

This translation process, known as featurization, represents the first critical step in applying ML to catalysis,<sup>7,8,9,10</sup> as it is the conceptual bridge between chemical intuition and data-driven modeling. For instance, if chemists want a ML model to quantify the steric hindrance of a target reactive site, they can use numerical descriptors such as buried volume ( $V_{\text{bur}}$ ).<sup>11</sup> This parameter estimates the accessibility of that site by measuring the fraction of a sphere occupied by neighboring atoms (Figure 1, top). Similarly, for electronic effects, chemists need to calculate features like electrostatic potential (ESP) or atomic charges. These values help capture whether a ring is electron-rich (i.e., with a NMe<sub>2</sub> group) or electron-poor (i.e., with a NO<sub>2</sub> group), transforming chemical intuition into machine-readable data that a model can interpret (Figure 1, bottom).



**Figure 1.** Examples of transforming concepts from chemical intuition into machine-readable descriptors.

Descriptors can capture the properties of an entire molecule (molecular descriptors) or specific atoms within it (atomic descriptors). In ML models for homogeneous catalysis, molecular descriptors are often complemented by atomic descriptors, as catalytic activity and selectivity frequently depend on the specific local environment around reactive sites.<sup>12</sup> Importantly, chemical intuition plays a central role in selecting which descriptors to generate, since the performance of ML models are strongly influenced by the relevance of the input features.<sup>13</sup> For example, a chemist experienced in Michael additions understands that the reaction outcome largely depends on the electrophilicity and steric accessibility of the carbon atom undergoing nucleophilic attack. Failing to include descriptors for this atom will likely reduce the accuracy of the resulting model. To capture these local effects in an ML framework, atomic descriptors for that carbon should contain electronic properties (i.e., partial charges, Fukui indices,<sup>14</sup> etc.) and steric parameters such as  $V_{\text{bur}}$  or solvent-accessible surface area (SASA).<sup>15</sup> Similarly, in metal-catalyzed reactions the metal center often dominates reactivity, and descriptors that capture the electronic and steric properties of that atom are essential for building accurate predictive models.

For non-specialists, the simplest entry point into descriptor generation is the use of online descriptor databases with quantum-mechanical (QM) features for large collections of ligands and catalysts. These tools offer researchers easy access to rich electronic and steric information without requiring custom calculations or coding skills. However, such databases typically cover only a limited range of structures and may be of limited use when the goal is to design new ligands. In practice, especially in catalysis, descriptors are often generated on-demand using QM methods. Density Functional Theory (DFT) and semiempirical approaches such as GFN2-xTB<sup>16</sup> are increasingly popular for calculating descriptors relevant to catalytic performance, including atomic charges, HOMO–LUMO energies, and Fukui indices, while steric parameters can be derived from the optimized quantum-mechanical structures.

This QM approach strategy is particularly valuable when experimental data is scarce, as is often the case in catalysis, where only a few dozen experiments may be available. However, when working with large-scale datasets, such as those generated in digital high-throughput screenings, QM calculations become computationally prohibitive. In these scenarios, alternative strategies such as molecular fingerprints from libraries like RDKit<sup>17</sup> or feature-learning methods based on graph neural networks (GNNs)<sup>18</sup> allow the rapid featurization of millions of molecules within reasonable computational times. These scalable methods are powerful when data is abundant, but they come with trade-offs: they often require greater programming expertise and generally lack the interpretability of physics-based descriptors.<sup>19</sup> Consequently, while fingerprints and GNNs are highly effective for large, high-volume datasets, they are rarely the first choice for the smaller, carefully curated datasets that are typical in catalysis research.

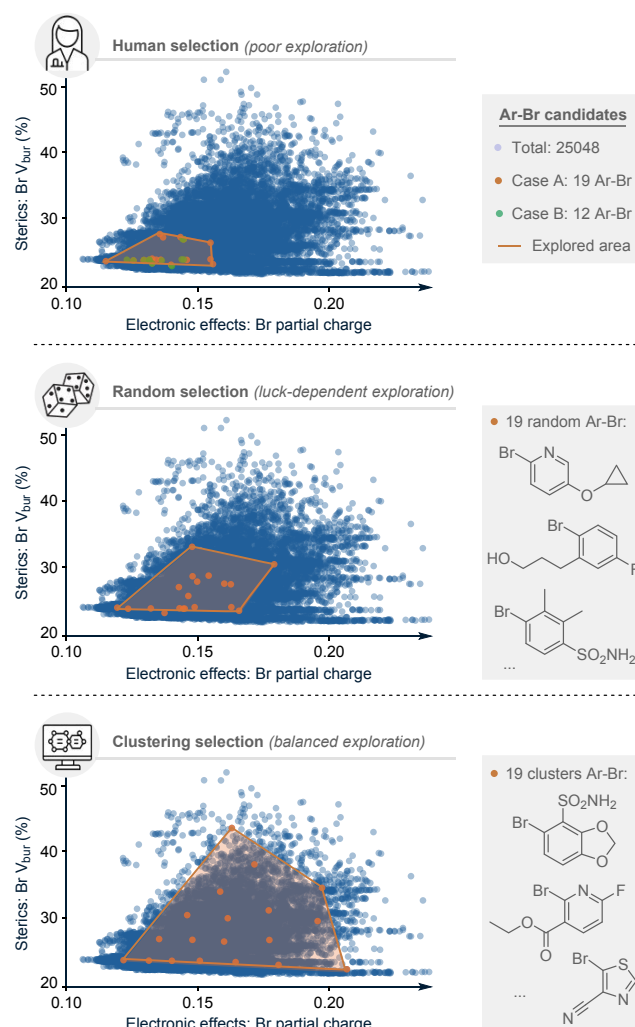
More information, implementation details, guidelines and considerations about descriptor generation are shown in Tables 1 and 2.

## SAMPLING SUBSTRATES AND CATALYSTS WITH DATA-DRIVEN CLUSTERING

One of the most common pitfalls of human intuition arises in selecting substrate scopes,<sup>20</sup> a standard practice used to evaluate

the generality of a catalytic method. In most articles, researchers tend to follow well-established patterns of chemical intuition to study how applicable the method is. They often explore electronic effects with a one-variable-at-a-time strategy such as substituting aromatics with a *para* electron-withdrawing group ( $\text{NO}_2$ ,  $\text{CF}_3$ ,  $\text{CN}$ ), hydrogen, or an electron-donating group ( $\text{OMe}$ ,  $\text{NMe}_2$ ). Similarly, steric effects are usually probed by replacing substituents such as H, Me,  $\text{'Pr}$ , and  $\text{'Bu}$  at a single position close to the reactive center. While this approach provides a general sense of how structural modifications influence catalytic activity, it represents an inefficient way of exploring chemical space and yields results that cover only a small fraction of the available diversity.

As an illustrative example, we downloaded and curated a set of 25048 commercially available bromoaryl ( $\text{Ar}-\text{Br}$ ) substrates from the Enamine chemical supplier, specifically from a collection designed for Pd-catalyzed cross-couplings.<sup>21</sup> We then mapped the chemical space<sup>22</sup> of substrates available from this vendor using two key descriptors: an electronic property (the partial charge on the leaving Br atom) and a steric parameter (the  $V_{\text{bur}}$  of Br).<sup>23</sup> The chemical space is represented with blue circles in the graphs of Figure 2. For comparison, we examined two reported metal-catalyzed cross-couplings of aromatic halides, case A<sup>24</sup> and case B,<sup>25</sup> which involved 19 and 12 substrates, respectively.



**Figure 2.** Chemical space exploration through human-guided, random, and data-driven sampling.

In both cases, the chosen molecules occupy only a small region of the vast and diverse chemical space (Figure 2, top, orange area), underscoring the limited scope and generality of the catalytic methods. Although these two cases serve purely as illustrative examples, similar trends likely extend across much of the cross-coupling literature, suggesting that relying solely on reported examples may provide an incomplete view of catalytic generality and limitations.

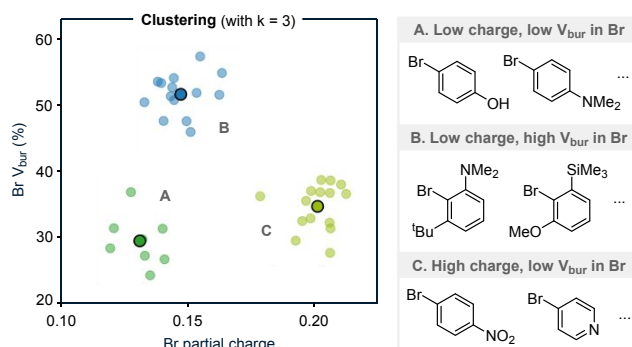
Based on this result one might argue that, once such a database of substrates is created, even random selection strategies could often explore a more diverse region of chemical space. In the example shown in Figure 2, middle, we picked substrates at random<sup>26</sup> and observed a considerable increase in the coverage of the explored region, although it still represented only a limited portion of the total space. A key drawback of this approach, however, is that random selections are inherently dependent on the chosen random seed and whether the sampled substrates cover a wide chemical space is essentially a matter of luck.

To overcome the inefficiencies of traditional substrate selection, unsupervised ML methods such as clustering are gaining traction in the chemistry community,<sup>27</sup> with notable applications from Sigman,<sup>28</sup> Doyle,<sup>29</sup> and Glorius,<sup>20</sup> among others. Clustering algorithms require only descriptors, without any pre-existing activity data, to partition chemical space into groups of compounds with similar properties. The resulting clusters reveal natural groupings without prior bias, and representative candidates can then be selected for targeted experimental validation.

An illustrative example using 38 Ar–Br structures and two properties ( $V_{\text{bur}}$  and the partial charge of Br) is shown in Figure 3. In this example, the molecules are separated into three clusters. Cluster A consists of phenyl rings with electron-donating groups in the *para* position relative to the Br atom and no proximal bulky substituents, corresponding to a low partial charge and low  $V_{\text{bur}}$  at the Br atom. Cluster B contains rings where the Br atom is flanked by bulky and electron-donating groups, appearing in the upper left region of the plot. Cluster C includes aromatics with *para* electron-withdrawing groups relative to the Br atom and no nearby steric repulsions, located in the lower right region of the graph.

Once the clusters are defined, the point closest to the centroid of each cluster can be selected as a representative molecule for sampling (Figure 3, dots with black borders). Using the larger-scale example from Figure 2, we generated 19 clusters using a *k*-means clustering algorithm and then selected the corresponding centroid molecules for comparison with the original 19 substrates from case A. Unlike intuition-driven selection, which is often guided by prior experience and limited to low-dimensional changes, clustering approaches objectively capture high-dimensional descriptor relationships and enable more efficient chemical exploration (Figure 2, bottom). By selecting a small set of substrates from distinct clusters, chemists can efficiently probe a more diverse and representative chemical space, ultimately gaining deeper insights into the generality and limitations of their catalytic methods.

In addition to guiding substrate selection, clustering can also be applied to explore catalyst diversity, a common strategy in homogeneous catalysis for optimizing reactivity and selectivity. Different groups have adopted this data-driven methodology, with notable examples from Pérez-Ramírez,<sup>30</sup> Jorner,<sup>31</sup> Ackermann,<sup>32</sup> and Denmark,<sup>33</sup> among others.



**Figure 3.** Example of clustering with three clusters.

Clustering has also been employed to discover catalysts by incorporating experimental results into generated clusters. For example, Schoenebeck and coworkers constructed a chemical space of ligands for the synthesis of palladium(I) dimers, which are challenging to stabilize.<sup>34</sup> Using a *k*-means algorithm, they divided ligand space into groups with related properties and subsequently incorporated stability data from five ligands. This analysis revealed that certain clusters were enriched in active ligands, whereas others contained inactive ones. This relatively simple strategy enabled the digital exploration of new phosphine ligands within the chemical space, and the prioritization of candidates from clusters containing stable representatives. Experimental validation ultimately led to the discovery of eight previously unexplored Pd dimers under conditions of very limited available data.

As a final remark for this section, it is important to note that more than three descriptors are typically used to featurize molecules and, consequently, to define chemical spaces. In such cases, the principal components obtained through principal component analysis (PCA)<sup>35</sup> can be employed to represent the chemical space in two- or three-dimensional plots. Plotting the sampling selection in these graphs helps verify that there is sufficient coverage of the chemical space. When using this approach, researchers should ensure that the selected principal components together capture a substantial portion of the dataset's variance (typically 60–70% or more).<sup>36</sup>

Guidelines and considerations relevant to this section are summarized in Tables 3 and 4 and should be carefully reviewed before attempting any clustering-based sampling.

## CATALYST DISCOVERY WITH SUPERVISED ML

Supervised learning refers to the development of predictive models from datasets that contain descriptors as the X matrix and one or more outcomes as the y matrix (i.e., reactivity, selectivity).<sup>37,38</sup> Broadly, the X matrix is processed by ML algorithms such as linear regression, random forests, and neural networks, which learn from the descriptor data to predict the y values. In catalysis, this learning of patterns across datasets makes it possible to map complex relationships between reaction parameters and performance metrics, enabling smarter experimental design and accelerating the discovery of optimal catalysts and conditions.<sup>3,39</sup>

Supervised learning problems in catalysis can generally be divided into regression and classification tasks. In regression, the objective is to predict numerical values within a continuous range, such as yields, rate constants, or enantioselectivity. In



contrast, classification problems aim to predict discrete outcomes, such as whether a catalyst is active or inactive, or to categorize performance levels (i.e., high vs low selectivity).<sup>40</sup> Notably, even small datasets containing as few as 18 reactions can deliver reliable predictions, provided that the chosen descriptors effectively capture the key chemical information.<sup>41</sup>

In both regression and classification problems, two main data-driven strategies are typically used for catalyst discovery: rational catalyst design enabled by explainable ML, and ML-based candidate prediction (Figure 4). These strategies are often integrated into iterative active learning cycles. In the first case, Strategy A employs a predictive model combined with SHAP feature analysis,<sup>42</sup> enabling chemists to gain mechanistic insights and prioritize the most informative candidates for testing through rational design. In the example from the figure, the SHAP analysis reveals that reduced steric hindrance and lower charge at Pd centers (Pd charge and Pd  $V_{\text{bur}}$ ) correlate with higher yields. This insight can help chemists to propose improved catalysts by incorporating groups that minimize these properties. A landmark example by Sigman and coworkers used multivariate linear regression to correlate steric and electronic descriptors of BINOL-derived phosphoric acids with experimental enantioselectivities, enabling accurate out-of-sample predictions (beyond the training set) and guiding ligand selection across diverse substrates.<sup>43</sup>

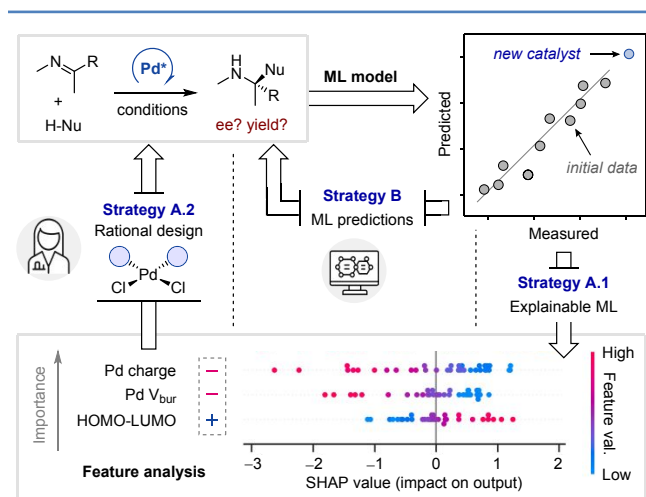
Alternatively, Strategy B focuses on discovering catalysts with minimal experimentation by relying solely on ML-guided suggestions. This strategy often employs Bayesian optimization (BO) to prioritize experiments, balancing options expected to work best (exploitation) with less certain choices that can provide new insights (exploration). In catalysis, this approach maximizes information gained per experiment while minimizing redundant trials. This discovery strategy is particularly popular for optimizing reaction conditions using descriptor matrices that include variables such as temperature, catalyst loading, and solvent. As an example, Doyle and co-workers demonstrated that BO identified high-yielding reaction conditions for a Pd-catalyzed arylation reaction faster than human chemists, highlighting its efficiency in navigating complex reaction landscapes.<sup>44</sup>

While this section highlights the advantages of using supervised ML and encourages its adoption, it is important for readers to understand that great care must be taken when developing and applying this technology. A double-edged situation is that many tools now allow chemists to build ML workflows and obtain predictions within minutes, regardless of their programming or data science expertise. Although this accessibility is crucial for the broader adoption of ML, the combination of limited expertise and the ease of model generation can be problematic, as it becomes relatively easy to overestimate the quality or the predictive power of our tools. For example, a researcher might train a model with an  $R^2$  of 0.95. These results could misleadingly suggest strong performance, when in fact the model is overfitted but went undetected because proper overfitting tests were skipped.

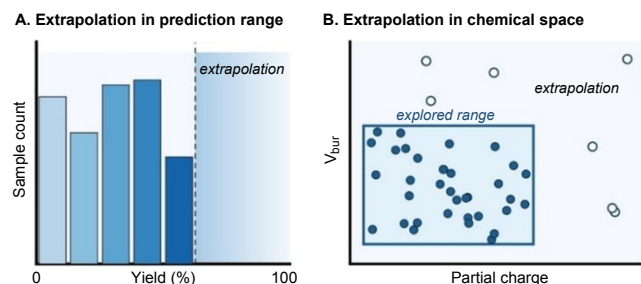
To avoid these pitfalls, researchers should evaluate models using multiple metrics (i.e.,  $R^2$ , RMSE, MAE, accuracy, F1 score, Matthews correlation coefficient) rather than relying on a single measure of performance. In addition, held-out test sets and  $k$ -fold cross-validation (preferably 5-fold CV) should be used to assess consistency and detect overfitting. Leave-one-out CV (LOOCV) can be applied to very small datasets but should be used with caution, as it is more brittle for detecting overfitting compared to 5-fold CV.<sup>45</sup> One example of an evaluation technique that combines these analyses with additional statistical tests (i.e., y-shuffle, y-mean, extrapolation) is the ROBERT score.<sup>46</sup> This metric evaluates models on a 10-point scale, considering three key aspects: predictive ability and overfitting, prediction uncertainty, and detection of spurious predictions.

Another essential consideration is the predictive scope of ML models, as they are often applied to predict outcomes for molecules beyond the range of their training data, where reliability drops sharply.<sup>47</sup> For instance, a researcher might develop an excellent predictor for the yields of substitution reactions in pyridines, but the same model may fail for pyrazines even if intuition suggests otherwise. In general, predictors should be treated with caution when applied to extrapolated regions outside their training sets, and experimental validation is strongly recommended in these cases. Examples of extrapolation include predicting yields higher than those observed in the training set (Figure 5A) and predicting outcomes for molecules substantially different from those used for training (Figure 5B).

Lastly, it is important to recognize that model accuracy depends largely on the quality of the training data. Studies on yield prediction consistently show a performance gap between models trained on high-throughput experimentation (HTE) datasets and those trained on electronic laboratory notebook (ELN) records.<sup>48</sup> HTE datasets typically follow strict, standardized protocols for reaction setup, analysis, and data logging, resulting in clean, balanced datasets with minimal missing information. By contrast, ELN-derived data are often accumulated over years by different researchers and tend to suffer from heterogeneous formats, incomplete metadata, and inconsistent experimental practices. For these reasons, another key consideration is that databases compiled from different manuscripts and patents often contain noisy yield values, which can severely undermine the accuracy of ML models.



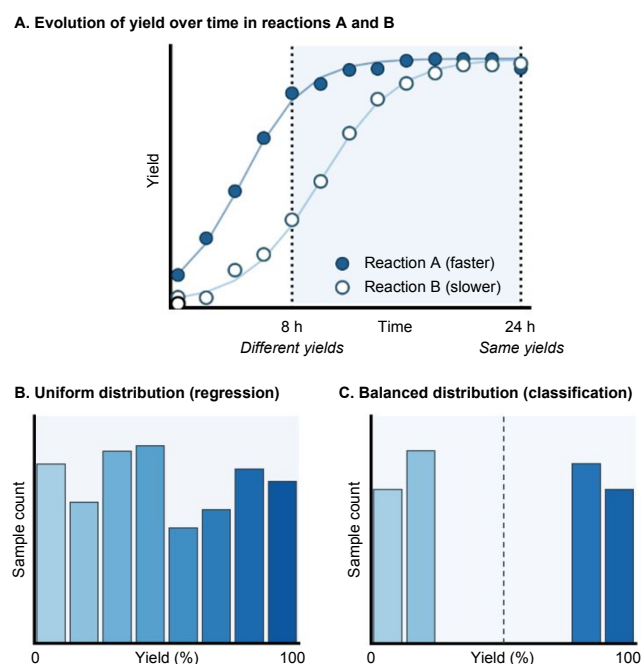
**Figure 4.** Comparison of different data-driven catalyst discovery strategies: rational design vs ML-based suggestions.



**Figure 5.** Different types of extrapolation in supervised ML.

In this context, chemists should, whenever possible, report crude yields (before purification) for comparability and maintain consistent conditions such as temperature, solvent, and sampling time when building or merging datasets. Reactions should also be performed in at least duplicate to ensure reproducibility and avoid spurious results. While selectivity might be less sensitive to some reaction parameters, caution is still recommended (i.e., diastereomeric ratios may change after column chromatography purification). Moreover, yield, conversion, and sometimes selectivity are strongly time-dependent (Figure 6A). It is therefore essential to select meaningful reaction times or to compare results across multiple time points.

For optimal performance, datasets should be balanced to minimize bias and improve model robustness. In regression, aim for a uniform distribution of target values (Figure 6B) so the model learns across low, medium, and high regions. In classification, ensure similar numbers of samples per class (Figure 6C) to avoid bias and loss of generalization. When data are highly skewed or concentrated around a single value, targeted data acquisition may be necessary to restore balance before modeling.<sup>49</sup>



**Figure 6.** Yield evolution over time in two reactions with different kinetics (A). Optimal distribution of target values for regression (B) and classification (C).

Further guidelines and considerations are summarized in Tables 5 and 6.

## ADVANCED ML APPLICATIONS IN CATALYSIS AND FUTURE DIRECTIONS

Despite the significant potential of ML to accelerate catalytic studies, its adoption within the chemistry community remains limited. A common reason for this is that implementing ML often requires time and effort to gain expertise in fields outside of chemistry (i.e., data science, programming). Closing this gap will require both education and rigor, including initiatives to embed hands-on digital chemistry modules into university curricula, offer workshops on ML applications, develop more user-friendly software, and adopt the Findable, Accessible, Interoperable, and Reusable (FAIR) principles.<sup>50</sup> In other cases, ML remains underutilized because researchers are skeptical. However, as with other disruptive technologies throughout history, this mistrust will likely fade as further advances emerge.

In the context of ML research, one of the most promising directions is the development of automated robotic platforms, which pave the way to popularize self-driving laboratories. Even though this is still a young field, several representative examples have already demonstrated success in catalytic reaction discovery, including work from the groups of Cooper,<sup>51</sup> Aspuru-Guzik,<sup>52</sup> and Noël,<sup>53</sup> among others.

Another emerging technology with growing influence in ML-driven catalysis is the use of large language models (LLMs). The creation of chatbot assistants holds great promise, as they can help chemists generate code, extract descriptors from published literature, and guide digital catalyst discovery through natural language conversations. Promising results have already been reported by the teams of Gomes,<sup>54</sup> White,<sup>55</sup> Laino,<sup>56</sup> and Schwaller,<sup>57</sup> among others.

For descriptor generation, a particularly exciting development is the emergence of machine learning potentials (MLPs), which enable simulations of catalytic systems with near-DFT accuracy at a fraction of the computational cost. While still at an early stage of adoption in catalysis, MLPs are poised to transform QM calculations and facilitate the routine exploration of complex reaction landscapes. Notable contributions in this area have been made by the groups of Isayev,<sup>58</sup> Dral,<sup>59</sup> Duarte,<sup>60</sup> and Wood and Zitnick,<sup>61</sup> among others.

Finally, an additional promising direction is ML-based inverse design, which aims to generate *in silico* catalysts by suggesting structural modifications through algorithms. These methods are often combined with filtering strategies to eliminate candidates that are too expensive or synthetically unfeasible. Active groups in this area include those of Balcells,<sup>62</sup> Jensen,<sup>63</sup> and Bhowmik,<sup>64</sup> among others.

## TABLES WITH FURTHER GUIDELINES AND CONSIDERATIONS

### Descriptor generation

**Table 1. Guidelines for generating descriptors.**

Step	Guidelines																
1. Descriptor generation	<p>Generate or collect relevant atomic and molecular descriptors to represent your system. A few options:</p> <ul style="list-style-type: none"><li>• Databases: KRAKEN,<sup>65</sup> ioChem-BD,<sup>66</sup> OSCAR,<sup>67</sup> and OMol25.<sup>68</sup></li><li>• Topological/MM descriptors and fingerprints: RDKit,<sup>17</sup> CDK,<sup>69</sup> and PaDELPy.<sup>70</sup></li><li>• QM descriptors: AQME,<sup>71</sup> ObeLiX,<sup>72</sup> Autoqchem,<sup>73</sup> and molli.<sup>74</sup></li><li>• Steric descriptors: MORFEUS,<sup>75</sup> DBSTEP,<sup>76</sup> and SambVca.<sup>77</sup></li><li>• ML-predicted properties: Chemprop<sup>78</sup> and MolPROP.<sup>79</sup></li></ul>																
2. Saving descriptors	<table><tr><th>Name</th><th>SMILES</th><th>Descriptors (X<sub>1</sub>, ..., X<sub>n</sub>)</th><th>Target value (y)</th></tr><tr><td>Mol<sub>1</sub></td><td>CO</td><td>X<sub>11</sub>, ..., X<sub>1n</sub></td><td>y<sub>1</sub></td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>Mol<sub>n</sub></td><td>CCO</td><td>X<sub>n1</sub>, ..., X<sub>nn</sub></td><td>y<sub>n</sub></td></tr></table> <p>Store generated descriptors in tabular (i.e., CSV) or JSON formats for easy sharing and reproducibility. Each entry should include key information such as SMILES strings (a compact way of encoding chemical structures),<sup>80</sup> descriptors (X values), and results (y values).</p>	Name	SMILES	Descriptors (X <sub>1</sub> , ..., X <sub>n</sub> )	Target value (y)	Mol <sub>1</sub>	CO	X <sub>11</sub> , ..., X <sub>1n</sub>	y <sub>1</sub>	...	...	...	...	Mol <sub>n</sub>	CCO	X <sub>n1</sub> , ..., X <sub>nn</sub>	y <sub>n</sub>
Name	SMILES	Descriptors (X <sub>1</sub> , ..., X <sub>n</sub> )	Target value (y)														
Mol <sub>1</sub>	CO	X <sub>11</sub> , ..., X <sub>1n</sub>	y <sub>1</sub>														
...	...	...	...														
Mol <sub>n</sub>	CCO	X <sub>n1</sub> , ..., X <sub>nn</sub>	y <sub>n</sub>														

**Table 2. Warnings and considerations for descriptor generation.**

Warning	Consideration
1. Conformational sampling	Many descriptors depend on molecular geometry, making conformational sampling essential. <sup>81</sup> Relying on a single conformer can produce misleading values if other low-energy conformations exist, a challenge especially common with flexible ligands and catalysts.
2. Relevant descriptors	While adding more descriptors can in principle improve predictive power, including irrelevant ones adds noise and increases the risk of overfitting, particularly in small datasets. <sup>82</sup> The most successful models focus on descriptors that capture the chemically meaningful variations in the data.

### Clustering

**Table 3. Guidelines for sampling with clustering.**

Step	Guidelines
1. Generate descriptors	Create meaningful atomic and molecular descriptors to define your chemical space (see the <i>Digitalization of Molecules</i> section).
2. Data curation	Data curation is essential in ML, and its quality directly influences the reliability of clustering. Thus, avoid using highly correlated descriptors or those that do not provide relevant molecular information, check for duplicated entries, outliers, etc. <sup>83,84</sup>
3. Run clustering algorithms and select samples	<p>Researchers might use available tools (i.e., Dedenser,<sup>85</sup> ALMOS,<sup>86</sup> Deep Clustering<sup>87</sup>) or web servers (i.e., iRaPCA,<sup>88</sup> Substrate Selection<sup>20</sup>) that are designed for chemists and often assist in descriptor generation as well.</p> <p>Chemists with basic Python knowledge may also create on-demand workflows using AI assistants. For example, the Jupyter Notebook used for clustering in Figure 2, bottom, was generated in under 5 minutes with ChatGPT.<sup>89</sup></p>

**Table 4. Warnings and considerations for clustering.**

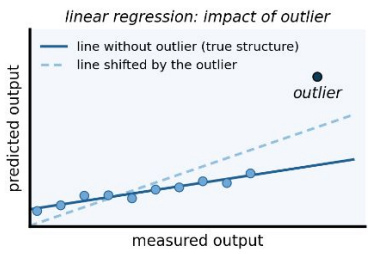
Warning	Consideration
1. Descriptor relevance	As noted in the previous section, it is essential to select descriptors relevant to the chemical problem. Generating a wide variety is easy, but irrelevant descriptors can dilute meaningful exploration.
2. Optimal number of clusters	The chosen number of clusters directly impacts the number of required experiments and, consequently, the resources, time, and waste involved. To determine the optimal number, techniques such as the elbow method and silhouette analysis can be used. <sup>34</sup>
3. Algorithm choice	Since clustering code is fast to create and execute, comparing different algorithms could maximize chemical space exploration while minimizing the number of required experiments. Common clustering algorithms in chemistry research include k-means, <sup>90</sup> HDBSCAN, <sup>91</sup> UMAP, <sup>92</sup> and t-SNE. <sup>93</sup>

### Supervised ML

**Table 5. Guidelines for supervised ML.**

Step	Guidelines
1. Generate descriptors	Create meaningful atomic and molecular descriptors (see the <i>Digitalization of Molecules</i> section).
2. Data curation	Data curation is essential in ML, and its quality directly influences the reliability of supervised models. Thus, avoid using highly correlated descriptors or those that do not provide relevant molecular information, check for repeated entries, outliers, etc. <sup>83,84</sup>
3. Choosing ML algorithms and automation of workflows	Use well-documented libraries such as scikit-learn, <sup>94</sup> XGBoost, <sup>95</sup> or Keras/TensorFlow. <sup>96,97</sup> They offer simple APIs for training, evaluation, and feature-importance analysis. It is helpful to compare the results obtained using different ML algorithms, such as linear regression, random forests and neural networks. For Bayesian optimization, tools such as EDBO+ <sup>98</sup> can be used.  Programs such as ROBERT <sup>99</sup> and DeepMol <sup>100</sup> are recommended to automate data curation, model selection, training, validation, and other ML tasks. These tools reduce the need for extensive manual coding and make ML more accessible to non-experts.

**Table 6. Warnings and considerations for supervised ML.**

Warning	Consideration
1. Overfitting	Monitor overfitting using CV (preferably 5-fold CV) and held-out test sets. High scores in the training set combined with substantially worse scores in CV or test set usually indicate memorization rather than learning. When possible, use composite metrics such as the ROBERT score to evaluate your models.
2. Outliers	 <p>Outliers are data points that deviate significantly from the general trend of the data. In supervised learning, they are often identified during or after model training by analyzing residuals and finding points that the model consistently predicts poorly. Outliers can bias model training, especially in small datasets, by forcing overfitting to rare or unrepresentative observations. Detecting and carefully reviewing such points is essential, as they often indicate errors in the data (i.e., incorrect yield assignment, incomplete geometry optimization) or molecules that differ significantly from the training set (i.e., predicting the aromaticity of a thiophene using a training set of pyridines). Outliers should not be removed without prior review.</p>



## ASSOCIATED CONTENT

**Data availability.** Raw data, instructions, and code used for descriptor generation and clustering are freely available on Zenodo (DOI: 10.5281/zenodo.17084325, <https://zenodo.org/records/17084325>).

## AUTHOR INFORMATION

### Corresponding Author

\* [jv.alegre@csic.es](mailto:jv.alegre@csic.es)

### Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally.

### Funding Sources

J.V.A.-R. acknowledges Gobierno de Aragón-Fondo Social Europeo (Research Groups E07\_23R), the State Research Agency

## REFERENCES

- Abraham, B. M.; Jyothirmmai, M. V.; Sinha, P.; Viñes, F.; Singh, J. K.; Illas, F. Catalysis in the digital age: Unlocking the power of data with machine learning. *WIREs Comput Mol Sci.* **2024**, *14*, e1730.
- de Araujo, L. G.; Vilcoq, L.; Fongarland, P.; Schuurman, Y. Recent developments in the use of machine learning in catalysis: A broad perspective with applications in kinetic. *Chem. Eng. J.* **2025**, *508*, 160872.
- Sanosa, N.; Dalmau, D.; Sampedro, D.; Alegre-Requena, J. V.; Funes-Ardoiz, I. Recent advances of machine learning applications in the development of experimental homogeneous catalysis *Artif. Intell. Chem.*, **2024**, *2*, 100068.
- Seavill, P. The future of digital chemistry. *Nat. Synth.* **2023**, *2*, 469–470.
- Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647.
- Dalmau, D.; Alegre-Requena, J. V. Integrating digital chemistry within the broader chemistry community. *Trends Chem.*, **2024**, *6*, 459–469.
- Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.
- Dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning. *Trends Chem.* **2021**, *3*, 96–110.
- Sigmund, L. M.; Assante, M.; Johansson, M. J.; Norrby, P.-O.; Jorner, K.; Kabeshov, M. Computational Tools for the Prediction of Site- and Regioselectivity of Organic Reactions. *Chem. Sci.* **2025**, *16*, 5383–5412.
- Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021**, *54*, 827–836.
- Escayola, S.; Bahri-Laleh, N.; Poater, A. %VBur index and steric maps: from predictive catalysis to machine learning. *Chem. Soc. Rev.* **2024**, *53*, 853–882.
- Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49*, 1292–1301.
- Seko, A.; Togo, A.; Tanaka, I. Descriptors for Machine Learning of Materials Data. In *Nanoinformatics*; Tanaka, I., Ed.; Springer Singapore: Singapore, 2018; pp 3–23.
- Fukui, K.; Yonezawa, T.; Shingu, H. A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons. *J. Chem. Phys.* **1952**, *20*, 722–725.
- Shrake, A.; Rupley, J. A. Environment and exposure to solvent of

of Spain (MCIN/ AEI/ 10.13039/501100011033/ FEDER, UE) for financial support (PID2022-140159NA-I00) and the European Union's Recovery and Resilience Facility-Next Generation in the framework of the General Invitation of the Spanish Government's public business entity Red.es to participate in talent attraction and retention programmes within Investment 4 of Component 19 of the Recovery, Transformation and Resilience Plan (MOMENTUM, MMT24-ISQCH-01).

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENT

J.V.A.-R. acknowledges the computing resources at the Galicia Supercomputing Center, CESGA, including access to the FinisTerae supercomputer and the Drago cluster facility of SGAI-CSIC. The authors also thank Chris Collison (Rochester Institute of Technology) for pre-reviewing the manuscript.

- protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **1973**, *79*, 351–371.
- Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
  - Landrum, G. RDKit: Open-source cheminformatics. <https://www.rdkit.org>
  - Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80.
  - Geemi P. Wellawatte, Heta A. Gandhi, Aditi Seshadri, Andrew D. White. A Perspective on Explanations of Molecular Prediction Models. *J. Chem. Theory Comput.* **2023**, *19*, 2149–2160.
  - Rana, D.; Pflüger, P. M.; Hölter, N. P.; Tan, G.; Glorius, F. Standardizing Substrate Selection: A Strategy toward Unbiased Evaluation of Reaction Generality. *ACS Cent. Sci.* **2024**, *10*, 899–906.
  - Ar–Br substrates were retrieved from the subset “Aryl Halides for Pd-Catalyzed Couplings” available at <https://enamine.net/building-blocks/functional-classes/aryl-halides> (accessed July 9, 2025). To construct a simplified chemical space of approximately 25,000 molecules, we applied the following filters: (1) exactly one Br atom, with zero Cl or I atoms, and (2) no counterions present.
  - Reymond, J. L. Chemical space as a unifying theme for chemistry. *J. Cheminform.* **2025**, *17*, 6.
  - Descriptors were calculated using AQME with GFN2-xTB, and clustering was performed using the *k*-means algorithm. These representations were generated for visualization purposes; alternative levels of theory or algorithms lead to the same conclusions regarding the limitations of human sampling.
  - Ackerman, L. K. G.; Lovell, M. M.; Weix, D. J. Multimetallic catalysed cross-coupling of aryl bromides with aryl triflates. *Nature* **2015**, *524*, 454–457.
  - Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **2018**, *362*, 670–674.
  - A random seed of 42 was used, which is a common choice in ML applications.
  - Talevi, A.; Bellera, C. L. Clustering of small molecules: new perspectives and their impact on natural product lead discovery. *Front. Nat. Produc.* **2024**, *3*, 1367537.
  - Samha, M. H.; Karas, L. J.; Vogt, D. B.; Odogwu, E. C.; Elward, J.; Crawford, J. M.; Steves, J. E.; Sigman, M. S. Predicting success in Cu-catalyzed C–N coupling reactions using data science. *Sci. Adv.* **2024**, *10*, eadn3478.
  - Kariofillis, S. K.; Jiang, S.; Zuranski, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *J. Am. Chem. Soc.* **2022**, *144*, 1045–1055.
  - Suvarna, M.; Zou, T.; Chong, S. H.; Ge, Y.; Martín, A. J.; Pérez-

- Ramírez, J. Active learning streamlines development of high performance catalysts for higher alcohol synthesis. *Nat. Commun.*, **2024**, *15*, 5844.
- <sup>31</sup> Schmid, S. P.; Schlosser, L.; Glorius, F.; Jorner, K. Catalysing (organo-)catalysis: Trends in the application of machine learning to enantioselective organocatalysis. *Beilstein J. Org. Chem.* **2024**, *20*, 2280–2304.
- <sup>32</sup> Hou, X.; Li, S.; Frey, J.; Hong, X.; Ackermann, L. Machine learning-guided yield optimization for palladaelectro-catalyzed annulation reaction. *Chem* **2024**, *10*, 2283–2294.
- <sup>33</sup> Olen, C. L.; Zahrt, A. F.; Reilly, S. W.; Schultz, D.; Emerson, K.; Candito, D.; Wang, X.; Strotman, N. A.; Denmark, S. E. Chemoinformatic Catalyst Selection Methods for the Optimization of Copper-Bis(oxazoline)-Mediated, Asymmetric, Vinylogous Mukaiyama Aldol Reactions. *ACS Catal.* **2024**, *14*, 2642–2655.
- <sup>34</sup> Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, L.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated dinuclear palladium catalyst identification through unsupervised machine learning. *Science*, **2021**, *374*, 1134–1140.
- <sup>35</sup> Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Ed. Psychol.* **1933**, *24*, 417–441.
- <sup>36</sup> Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A* **2016**, *374*, 20150202.
- <sup>37</sup> Żurański, A. M.; Martínez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54*, 1856–1865.
- <sup>38</sup> Oliveira, J. C. A.; Frey, J.; Zhang, S.-Q.; Xu, L.-C.; Li, X.; Li, S.-W.; Hong, X.; Ackermann, L. When Machine Learning Meets Molecular Synthesis. *Trends Chem.* **2022**, *4*, 863–885.
- <sup>39</sup> Morán-González, L.; Burnage, A. L.; Nova, A.; Balcells, D. AI Approaches to Homogeneous Catalysis with Transition Metal Complexes. *ACS Catal.* **2025**, *15*, 9089–9105.
- <sup>40</sup> Wong, Y.-P.; Jung, H.-J.; Lin, S.; Shammami, M. A.; Roshandel, H.; Dodge, H. M.; Chapp, S. M.; Ruiz De Castilla, L. C.; Wang, D.; Do, L. H.; Liu, C.; Miller, A. J. M.; Diaconescu, P. L. Using Classifiers To Predict Catalyst Design for Polyketone Microstructure. *J. Am. Chem. Soc.* **2025**, *147*, 3913–3918.
- <sup>41</sup> Dalmau, D.; Sigman, M. S.; Alegre-Requena, J. V. Machine learning workflows beyond linear models in low-data regimes. *Chem. Sci.* **2025**, *16*, 8555–8560.
- <sup>42</sup> Shapley, L. S. in *The Shapley Value*, ed. A. E. Roth, Cambridge University Press, 1st edn., 1988, pp. 31–40.
- <sup>43</sup> Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343–348.
- <sup>44</sup> Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, **2021**, *590*, 89–96.
- <sup>45</sup> Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **2018**, *180*, 68–77.
- <sup>46</sup> <https://robert.readthedocs.io/en/latest/Report/score.html>
- <sup>47</sup> Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of Prediction Confidence and Domain Extrapolation of Two Structure–Activity Relationship Models for Predicting Estrogen Receptor Binding Activity. *Environ. Health Perspect.* **2004**, *112*, 1249–1254.
- <sup>48</sup> Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Żurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the Use of Real-World Datasets for Reaction Yield Prediction: Performance Differences Between HTE and ELN Data. *Chem. Sci.* **2023**, *14*, 4997–5005.
- <sup>49</sup> Haas, B. C.; Kalyani, D.; Sigman, M. S. Applying Statistical Modeling Strategies to Sparse Datasets in Synthetic Chemistry. *Sci. Adv.* **2025**, *11*, ead3013.
- <sup>50</sup> Wilkinson, M.; Dumontier, M.; Aalbersberg, I.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
- <sup>51</sup> Dai, T.; Vijayakrishnan, S.; Szczepiński, F. T.; Ayme, J.-F.; Simaei, E.; Fellowes, T.; Clowes, R.; Kotopantov, L.; Shields, C. E.; Zhou, Z.; Ward, J. W.; Cooper, A. I. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **2024**, *635*, 890–897.
- <sup>52</sup> Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-
- Guzik, A. Materials Acceleration Platforms: On the way to autonomous experimentation. *Curr. Opin. Green Sus. Chem.* **2020**, *25*, 100370.
- <sup>53</sup> Slattery, A.; Wen, Z.; Tenblad, P.; Sanjosé-Orduna, J.; Pintossi, D.; Den Hartog, T.; Noël, T. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science* **2024**, *383*, eadj1817.
- <sup>54</sup> Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language model. *Nature* **2023**, *624*, 570–578.
- <sup>55</sup> Ramos, M. C.; Collison, C. J.; White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **2025**, *16*, 2514–2572.
- <sup>56</sup> Teukam, Y. G. N.; Dassi, L. K.; Manica, M.; Probst, D.; Schwaller, P.; Laino, T. Language models can identify enzymatic binding sites in protein sequences. *Comput. Struct. Biotech. J.* **2024**, *23*, 1929–1937.
- <sup>57</sup> Bran, M. A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. S.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525–535.
- <sup>58</sup> Anstine, D. M.; Zubatyuk, R.; Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **2025**, *16*, 10228–10244.
- <sup>59</sup> Alavi, S. F.; Chen, Y.; Hou, Y.-F.; Ge, F.; Zheng, P.; Dral, P. O. ANI-1ccx-gelu Universal Interatomic Potential and Its Fine-Tuning: Toward Accurate and Efficient Anharmonic Vibrational Frequencies. *J. Phys. Chem. Lett.* **2025**, *16*, 483–493.
- <sup>60</sup> Zhang, H.; Juraskova, V.; Duarte, F. Modelling chemical processes in explicit solvents with machine learning potentials. *Nat. Commun.* **2024**, *15*, 6114.
- <sup>61</sup> Wood, B. M.; Dzamba, M.; Fu, X.; Gao, M.; Shuaibi, M.; Barroso-Luque, L.; Abdelmaqsood, K.; Gharakhanyan, V.; Kitchin, J. R.; Levine, D. S.; Michel, K.; Sriram, A.; Cohen, T.; Das, A.; Rizvi, A.; Sahoo, S. J.; Ulissi, Z. W.; Zitnick, C. L. UMA: A Family of Universal Models for Atoms. *arXiv* **2025**, 2506.23971.
- <sup>62</sup> Strandgaard, M.; Linjordet, T.; Kneiding, H.; Burnage, A. L.; Nova, A.; Jensen, J. H.; Balcells, D. A Deep Generative Model for the Inverse Design of Transition Metal Ligands and Complexes. *JACS Au* **2025**, *5*, 2294–2308.
- <sup>63</sup> Seumer, J.; Jensen, J. H. Beyond predefined ligand libraries: a genetic algorithm approach for de novo discovery of catalysts for the Suzuki coupling reactions. *PeerJ Phys. Chem.* **2025**, *7*, e34.
- <sup>64</sup> Cornet, F.; Benediktsson, B.; Hastrup, B.; Schmidt, M. N.; Bhowmik, A. OM-Diff: inverse-design of organometallic catalysts with guided equivariant denoising diffusion. *Dig. Discov.* **2024**, *3*, 1793–1811.
- <sup>65</sup> Gensch, T.; Gomes, G. D. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.
- <sup>66</sup> Álvarez-Moreno, M.; de Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the computational chemistry big data problem: the ioChem-BD platform. *J. Chem. Inf. Model.* **2015**, *55*, 95–103.
- <sup>67</sup> Gallarati, S.; van Gerwen, P.; Laplaza, R.; Vela, S.; Fabrizio, A.; Corminboeuf, C. OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts. *Chem. Sci.* **2022**, *13*, 13782–13794.
- <sup>68</sup> Levine, D. S.; Shuaibi, M.; Spotte-Smith, E. W. C.; Taylor, M. G.; Hasyim, M. R.; Michel, K.; Batatia, I.; Csányi, G.; Dzamba, M.; Eastman, P.; Frey, N. C.; Fu, X.; Gharakhanyan, V.; Krishnapriyan, A. S.; Rackers, J. A.; Raja, S.; Rizvi, A.; Rosen, A. S.; Ulissi, Z.; Vargas, S.; Zitnick, C. L.; Blau, S. M.; Wood, B. M. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. *arXiv* **2025**, 2505.08762.
- <sup>69</sup> Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliakova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*, 33.
- <sup>70</sup> Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.
- <sup>71</sup> Alegre-Requena, J. V.; Sowndarya, S. V.; Shree; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S. AQME: Automated Quantum Mechanical

Environments for Researchers and Educators. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1663.

<sup>72</sup> Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. Paving the road towards automated homogeneous catalyst design. *ChemPlusChem* **2024**, *89*, e202300702.

<sup>73</sup> Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: An Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276–1284.

<sup>74</sup> Shved, A. S.; Ocampo, B. E.; Burlova, E. S.; Olen, C. L.; Rinehart, N. I.; Denmark, S. E. molli: A General-Purpose Python Toolkit for Combinatorial Small Molecule Library Generation, Manipulation, and Feature Extraction. *J. Chem. Inf. Model.* **2024**, *64*, 8083–8090.

<sup>75</sup> Jorner, J. <https://github.com/digital-chemistry-laboratory/morfeus>

<sup>76</sup> Luchini, G.; Patterson, T.; Paton, R. S. DBSTEP: DFT Based Steric Parameters. 2022, DOI: 10.5281/zenodo.4702097.

<sup>77</sup> Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics* **2016**, *35*, 2286–2293.

<sup>78</sup> Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9–17.

<sup>79</sup> Rollins, Z. A.; Cheng, A. C.; Metwally, E. MolPROP: Molecular Property prediction with multimodal language and graph fusion. *J. Cheminform.* **2024**, *16*, 56.

<sup>80</sup> Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

<sup>81</sup> Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. Impact of Model Selection and Conformational Effects on the Descriptors for *In Silico* Screening Campaigns: A Case Study of Rh-Catalyzed Acrylate Hydrogenation. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

<sup>82</sup> Kuncheva, L. I.; Matthews, C. E.; Arnaiz-González, Á.; Rodríguez, J. J. Feature Selection from High-Dimensional Data with Very Low Sample Size: A Cautionary Tale. *arXiv* **2020**, 2008.12025.

<sup>83</sup> Comesana, A. E.; Huntington, T. T.; Scown, C. D.; Niemeyer, K. E.; Rapp, V. H. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. *Fuel* **2022**, *321*, 123836.

<sup>84</sup> Xerxa, E.; Vogt, M.; Bajorath, J. Influence of Data Curation and Confidence Levels on Compound Predictions Using Machine Learning Models. *J. Chem. Inf. Model.* **2024**, *64*, 9341–9349.

<sup>85</sup> Beck, A. G.; Fine, J.; Lam, Y.-H.; Sherer, E. C.; Regalado, E. L.; Aggarwal, P. Dedenser: A Python Package for Clustering and Downsampling Chemical Libraries. *J. Chem. Inf. Model.* **2025**, *65*, 1053–1060.

<sup>86</sup> Martinez-Fernandez, M. <https://github.com/MiguelMartzFdez/almos>

<sup>87</sup> Hadipour, H.; Liu, C.; Davis, R.; Cardona, S. T.; Hu, P. Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. *BMC Bioinformatics* **2022**, *23*, 132.

<sup>88</sup> Gori, D. N. P.; Llanos, M. A.; Bellera, C. L.; Talevi, A.; Alberca, L. N. iRaPCA and SOMoC: Development and Validation of Web Applications for New Approaches for the Clustering of Small Molecules. *J. Chem. Inf. Model.* **2022**, *62*, 2987–2998.

<sup>89</sup> GPT-5 model (accessed August 7, 2025). The prompts used and the resulting code are available in the Zenodo repository containing the supplementary data for this Viewpoint.

<sup>90</sup> MacQueen, J. Some methods for classification and analysis of multivariate observations. *Berkeley Symp. Math. Statist. Prob.* **1967**, 281–297.

<sup>91</sup> Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, **2013**, 7819. Springer, Berlin, Heidelberg.

<sup>92</sup> McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, 1802.03426.

<sup>93</sup> van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

<sup>94</sup> Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.;

Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

<sup>95</sup> Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco California USA, 2016; pp 785–794.

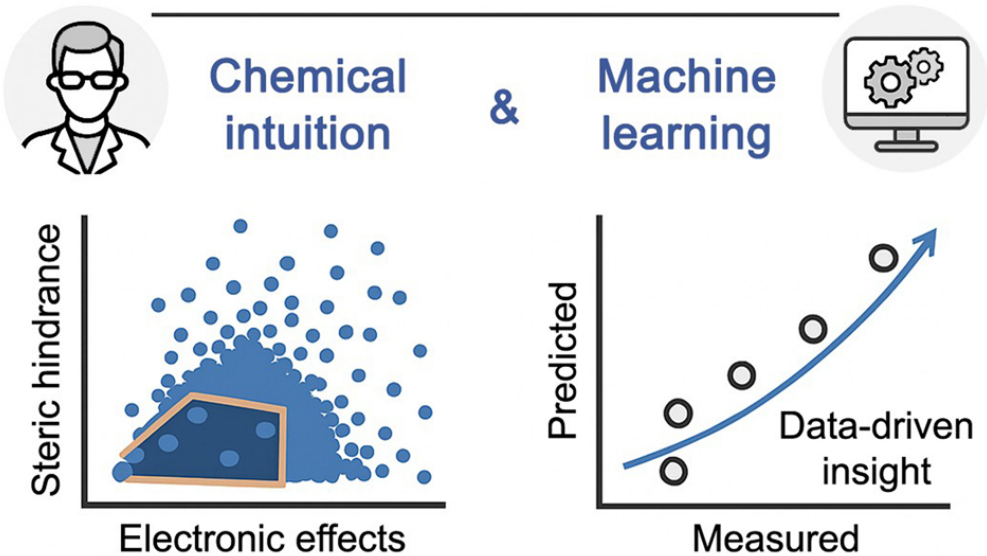
<sup>96</sup> Chollet, F. and others. Keras, **2015**. <https://keras.io>.

<sup>97</sup> Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, 1603.04467.

<sup>98</sup> Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144*, 19999–20007.

<sup>99</sup> Dalmau, D.; Alegre-Requena, J. V. ROBERT: Bridging the Gap Between Machine Learning and Chemistry. *WIREs Comput Mol Sci* **2024**, *14*, e1733.

<sup>100</sup> Correia, J.; Capela, J.; Rocha, M. Deepmol: An Automated Machine and Deep Learning Framework for Computational Chemistry. *J. Cheminform.* **2024**, *16*, 136.



80x45mm (300 x 300 DPI)