ACS Catalysis

pubs.acs.org/acscatalysis

Research Article

Cost-Effective Quantum-Mechanical Workflows for Molecular Machine Learning

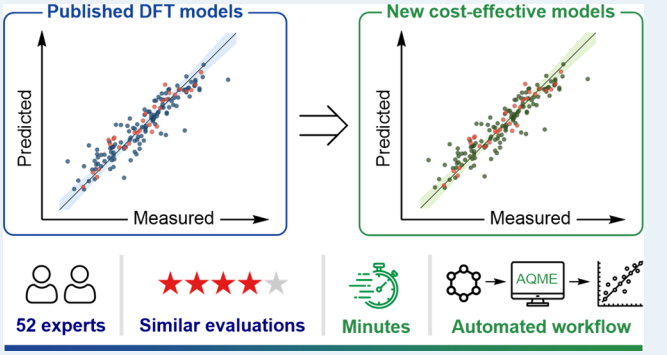
David Dalmau, Lauriane Jacot-Descombes, Adarsh Kalikadien, Brenda Manzanilla, Evgeny A. Pidko, Kjell Jorner, Matthew S. Sigman, and Juan V. Alegre-Requena*

Cite This: *ACS Catal.* 2026, 16, 12565−12574

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The current digital revolution has driven a rapid adoption of machine learning (ML) to support decision-making in chemistry. In this context, many studies now incorporate ML models based on quantum-mechanical (QM) descriptors to accelerate the discovery of candidates across a wide range of chemical problems. While density functional theory (DFT) has traditionally served as the primary source of such descriptors, recent advances in cost-effective QM methods offer an opportunity to reduce computational cost and facilitate faster integration of ML into chemical workflows. Herein, we present an automated and user-friendly workflow that generates 39 electronic and steric cost-effective descriptors and their integration into ML models. These models are broadly applicable to fields involving finite molecular systems, such as homogeneous catalysis and drug discovery. We evaluated the acceptance of these cost-effective descriptors in ML-driven catalysis through a blinded survey of 52 participants, which indicated that experts consider the resulting ML models to be as reliable and interpretable as publication-quality DFT-based models. Finally, we introduced a fully automated protocol that generates descriptors and ML predictors from SMILES strings, enabling rapid predictor development accessible to the broader chemistry community. This cost-effective framework holds the potential to accelerate the integration of ML into everyday chemical research and contribute to a more inclusive digital transformation of the field.



KEYWORDS: *quantum-mechanical descriptors, machine learning, cheminformatics, automation, cost-effective workflows*

INTRODUCTION

The integration of machine learning (ML) into chemical discovery has transformed the field, enabling the prediction of reaction outcomes, properties, and optimal conditions that would otherwise require extensive experimentation.[1−5] This digital technology provides frameworks for learning structure−reactivity relationships from experimental and computational data, supporting more informed experimental design and accelerating catalyst development.[6−8] In this context, ML has found numerous applications in chemistry, including substrate selection,[9−11] discovery of ligands, catalysts, and reactions,[12−14] reaction condition optimization,[15,16] and property prediction,[17−19] among others.

A critical component of ML modeling is the choice of molecular descriptors, which encode structural, electronic, and steric information into features that algorithms can learn from.[20,21] Quantum-mechanical (QM) descriptors are often used to capture physically meaningful properties that correlate with reactivity,[22−26] with density functional theory (DFT) currently standing as the preferred framework for the generation of QM descriptors due to its balance of accuracy and general applicability.[27] However, the computational cost of DFT calculations limits their scalability for large datasets, iterative workflows, and rapid screening of extensive candidate spaces, representing a key bottleneck preventing chemists from a wider implementation of ML.

As a cost-effective alternative to DFT, semiempirical QM methods such as the GFN-xTB family have gained popularity, providing sufficiently accurate approximations of the electronic structure at significantly lower computational expense.[28,29] Despite these advantages, semiempirical QM approaches remain underutilized in many ML workflows within the chemistry community, largely due to perceived limitations of lower-level methods. However, for many classes of molecules commonly encountered in ML studies, such cost-effective descriptors have

Received: April 2, 2026
Revised: May 9, 2026
Accepted: May 11, 2026
Published: May 26, 2026

proven to be useful. For example, recent work has shown that xTB-derived descriptors can capture relevant electronic trends and correlate with physical organic parameters in contexts where DFT is typically used.[30−32] Similarly, xTB descriptors have been employed to train ML models with performance comparable to their DFT-based analogues.[33,34]

In this study, we examine the practical use of cost-effective QM descriptors in ML-driven chemistry. Using electronic and steric descriptors generated with the xTB and MORFEUS programs,[35] we train ML models of representative organic and organometallic finite systems reported in recent literature. To assess how such models are evaluated in practice, we conducted a blinded survey of experts in chemical ML, who compared models built using cost-effective descriptors with publication-quality models based on DFT (Figure 1). The survey provides direct insight into perceived reliability, interpretability, and model preference from an expert user perspective. Finally, we introduce an automated AQME[36]-ROBERT[37] workflow that enables end-to-end generation of QM descriptors and ML models from SMILES strings, thereby facilitating the rapid and reproducible deployment of ML methods in chemical research.[38]

RESULTS AND DISCUSSION

Automated Descriptor Generation

The performance and practical utility of ML models in chemistry are strongly influenced by the choice of molecular descriptors. QM features typically include a limited set of properties directly extracted from calculations, such as atomic charges, frontier orbital energies, and dipole moments.[39] These are sometimes complemented by conceptual DFT (cDFT) descriptors, which relate electron density to chemical reactivity (i.e., Fukui indices or ionization potentials),[40,41] as well as steric descriptors, such as buried volumes (V_{bur}).[42] Despite the wide range of available QM descriptor types, many studies rely on relatively small and heterogeneous subsets, which potentially leaves a fraction of chemically relevant information unexplored. For example, while popular choices such as Fukui indices, HOMO/LUMO energies, atomic charges, sterimol parameters,[43] and V_{bur} show a notable presence in the catalysis literature,[23,44−47] other complementary features are often overlooked or not combined.

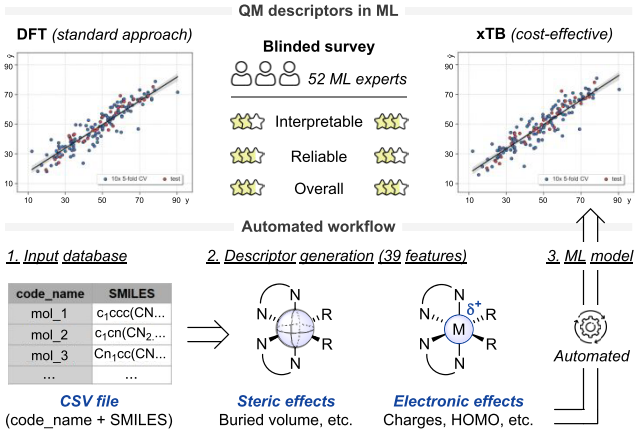


Figure 1. Overview of the descriptor generation workflow and its application in ML modeling.

To address this limitation, we developed a new automated AQME workflow coupled with xTB and MORFEUS to calculate a collection of 39 descriptors (the "interpret" set), comprising 21 molecular and 18 atomic descriptors (Figure 2A). This set includes electronic features such as atomic charges, Fukui indices, HOMO/LUMO energies, hardness/softness, and nucleophilicity/electrophilicity, among others, as well as steric parameters such as V_{bur}, solvent-accessible surface areas (SASA), and pyramidalization. The complete list and definitions of the descriptors are provided in Table S1. This set is the standard choice for ML modeling, as it provides broad, potentially relevant, and chemically interpretable molecular representations.

These 39 descriptors combine different interpretable features calculated with xTB and MORFEUS, together with additional properties identified as relevant in the cDFT literature. As new descriptor definitions and software updates become available, this list will likely be expanded to include additional QM-based descriptors. Alternatively, the workflow generates a larger descriptor dataset ("full"), comprising 238 descriptors that combine xTB, MORFEUS, and RDKit[48] features. However, this expanded set is not discussed in this work[49] since its primary aim is to maximize predictive power, often at the expense of interpretability.

The workflow first performs a conformer sampling based on RDKit, while incorporating internal AQME protocols to overcome known limitations. Although more exhaustive conformational exploration can be achieved with methods such as GOAT[50] and CREST,[51] RDKit provides a practical balance between computational efficiency and conformational coverage for high-throughput molecular modeling workflows. In this context, the AQME implementation extends the applicability of the RDKit-based search beyond standard organic molecules through internal filters that enable the treatment of organometallic systems, allows control over molecular geometry (i.e., tetrahedral vs. square planar complexes), and provides options to manage *cis*−*trans* isomerism.[49] This is consistent with recent efforts to expand RDKit-compatible SMILES representations to transition metal complexes, further supporting SMILES-driven workflows for organometallic ML.[52] The resulting geometries are optimized using the GFN2-xTB method, and descriptors are subsequently calculated with MORFEUS for steric properties and with GFN2-xTB and PTB[53] for electronic features through an interface in MORFEUS.

While xTB may exhibit significantly larger absolute errors than DFT due to its lower level of theory, previous studies have shown that xTB-based descriptors can effectively capture the relative trends observed with DFT.[54] This is particularly relevant because the descriptors generated within the proposed AQME workflow are intended for ML modeling, where relative trends are often more important than absolute values. With this in mind, we performed a benchmarking study using steric (V_{bur}) and electronic (HOMO energy, HOMO−LUMO gap, HL gap, and dipole module, μ) descriptors. The benchmark included different neutral and charged organic and organometallic compounds,[37,55−57] comparing the same conformer optimized and featured with either xTB or DFT (Figure 2B).

In this test, a reasonable R^2 threshold of 0.7 (gray line) was adopted to identify meaningful correlations. For V_{bur} (pale blue bars), R^2 values were consistently above this threshold and close to 1.0 for **1**−**5**. This can be expected, as V_{bur} depends only on the geometry and the GFN2-xTB method has been parameterized to reproduce DFT structures. For the electronic descriptors calculated with PTB (top blue bars), only one correlation

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565−12574

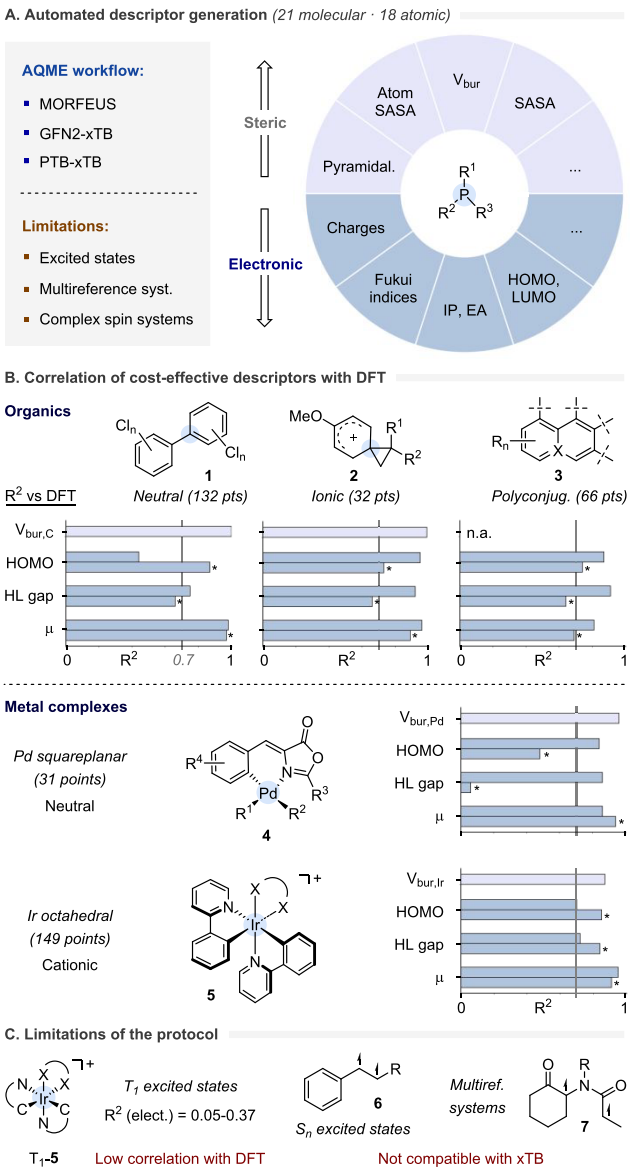


Figure 2. (A) Descriptors generated within the new AQME workflow. (B) R^2 comparison of the generated descriptors against DFT results, highlighting known limitations. V_{bur} was calculated with MORFEUS (pale blue bars). The HOMO, HL gap, and μ were calculated with PTB (top blue bars). Asterisks (*) represent values calculated with GFN2-xTB (bottom blue bars). (C) Limitations of the method.

fell below 0.7, corresponding to the HOMO of family **1**. In contrast, GFN2-xTB generally showed lower correlation with DFT compared to PTB, with particularly poor performance in Pd complexes **4** (i.e., an R^2 of 0.06 for the HL gap). Similar weak correlations with GFN2-xTB have also been observed previously in other Pd complexes.[54] For this reason, PTB is used as the default method within the AQME workflow for all electronic properties compatible with this xTB-based protocol.

Overall, these results demonstrate that the proposed descriptors effectively capture DFT trends across diverse chemical systems, supporting their use as cost-effective alternatives for general ML modeling (Figure 2B and Table S2). Nevertheless, the methodology remains subject to intrinsic limitations of xTB, and certain chemical systems may yield unreliable results and should therefore be treated with caution (Figure 2C). These include excited states (T_1-**5** and **6**) and multireference systems (**7**),[31,58] as well as molecules containing ionic or highly polar interactions and some transition metal complexes, for which GFN*n*-xTB methods may overstabilize strained motifs or yield distorted, qualitatively incorrect geometries.[59] As upcoming xTB methods such as g-xTB become more established, their integration into the workflow may help alleviate some of these limitations while extending its applicability to additional elements beyond those currently covered by GFN*n*-xTB methods.

Cost-Effective Descriptors in Unsupervised ML Problems

To demonstrate the practical utility of the generated descriptors in ML, we compared cost-effective and DFT-derived descriptors in an unsupervised task using a dataset of 245 monophosphine ligands (Figure 3).[60] For the DFT reference, we selected 12 commonly used ligand descriptors from the Kraken library[61] computed on DFT-optimized geometries and Boltzmann-weighted over multiple conformers, including steric (i.e., P pyramidalization) and electronic (i.e., the HL gap, μ, P partial charge, hardness, etc.) properties. In parallel, the corresponding cost-effective descriptor set was generated fully automatically using the AQME workflow, starting from SMILES strings and coupling conformational sampling with xTB-based descriptor generation.

Both descriptor sets were independently projected using UMAP dimensionality reduction[62] and clustered into three groups using HDBSCAN.[63−65] In the clustering analysis (Figure 3), the similarity between the resulting cluster assignments was quantified using the normalized mutual information (NMI) score, where values of 0 and 1 indicate no agreement and perfect agreement, respectively.[66,67] We obtained an NMI score of 0.84, indicating very good agreement between the cluster assignments obtained from cost-effective and DFT-derived descriptors.

Similarly, 237 out of 245 ligands received matching assignments, indicating that both types of descriptors lead to practically identical clustering results. Among the eight ligands with nonmatching cluster assignments, two exhibit the largest discrepancies between AQME- and Kraken-derived descriptor values, explaining their different clustering (Figures S3 and S4). For the remaining six, the origin of the mismatch is less clear and appears to arise from the cumulative effect of moderate deviations across multiple descriptors. Overall, the strong agreement between the two methods in this unsupervised test supports that the proposed cost-effective descriptors capture

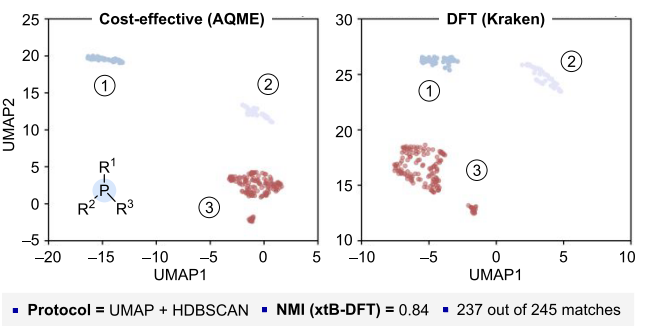


Figure 3. UMAP-HDBSCAN protocol using 245 phosphines with 12 xTB- and DFT-based descriptors. The three resulting clusters are labeled in the graphs and shown in different colors.

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565−12574

similar underlying features as DFT, despite their significantly lower computational cost.

Supervised ML and Expert Survey

Motivated by the successful results observed in the previous unsupervised ML task, we next focused on supervised examples relevant to homogeneous catalysis. Five recent case studies were selected to evaluate whether cost-effective descriptors can compete with publication-quality DFT-based models (Figure 4). To increase the difficulty of the benchmark, the selected examples span a diverse range of systems, including organic and organometallic species, different spin multiplicities, combination of descriptors from different molecules, and varying molecular sizes. Specifically, cases A[68] and B[45] involve radical Cu complexes and diamagnetic Ni species, cases C[46] and D[69] combine features from multiple organic molecules, including organocatalysts with 70-90 atoms, and case E[47] involves boryl radicals. The selected datasets encompass the typical sizes commonly encountered in related literature,[70] ranging from 18 to 153 entries, and included both regression (A–D) and classification (E) problems.

Cost-effective descriptors were generated from SMILES strings using the automated AQME workflow, as explained in the previous section. In examples A and B, the structures of relevant metal complex intermediates were used as inputs, while in example E, the starting boryl radicals were employed. In examples C and D, descriptors derived from three or two reaction components, respectively, were combined.

Manual feature selection of descriptors was then carried out to achieve a balance between model reliability and interpretability. Following the strategy adopted in the original DFT studies, descriptor combinations were systematically evaluated up to the maximum number used in the corresponding DFT models, and the optimal models were selected based on their performance metrics, with preference given to those providing greater interpretability. Although researchers may use the fully automated AQME-ROBERT workflow (see the next section), we adopted a manual curation strategy in this benchmark because it often yields superior models for small datasets and for applications where interpretability is important.

The selected cost-effective descriptors had a similar electronic or steric nature to those used in the original DFT-based models (Figures S6–S10, top). For instance, in example B, the original model used three electronic features (charges on the N and C atoms and molecular polarizability), whereas the cost-effective model used two electronic features (nucleophilicity of the N atom and polarizability of the C atom). The interpretability of both types of approaches is assessed below.

Subsequently, ML models were trained using ROBERT, which enables robust and systematic comparisons by (i) avoiding human bias in data splitting and (ii) employing both 10-times repeated 5-fold cross-validation[71,72] (10× 5-fold CV) and external test sets to prevent overfitting and misleading performance. The ROBERT score was also used to evaluate model quality. This metric is a composite score out of 10 that accounts for multiple aspects of model performance, including overfitting, interpolation and extrapolation behavior, predictive accuracy, and robustness (Figure S5).[71,73] In all cases, we ensured a consistent

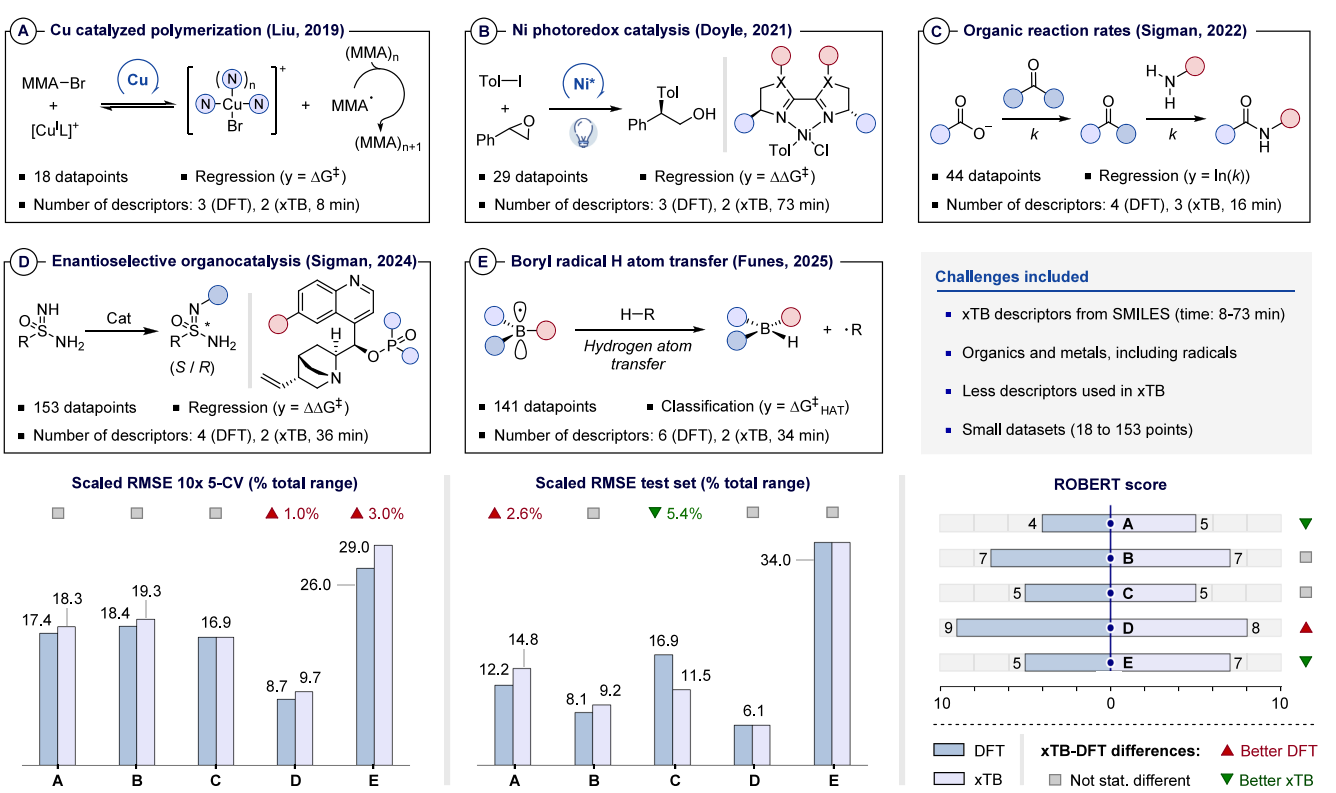


Figure 4. Performance of cost-effective and DFT-based ML models, showing scaled root-mean-square error (RMSE) for cases A–D, accuracy error for case E, and ROBERT scores. Gray squares denote models whose error performances are not statistically different (based on *t*-tests and Wilcoxon signed-rank tests) or have the same ROBERT score. Triangles indicate better performance with DFT (red) or xTB (green). Workflow followed: SMILES to descriptors with AQME, then manual feature selection, and lastly ML modeling with ROBERT.

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565–12574

comparison between cost-effective and DFT-based models by using the same class of algorithm (i.e., linear or nonlinear, as appropriate, Figures S6–S10).

To assess the performance of the cost-effective models, we compared them with their corresponding published DFT-based models using the same datasets and descriptors reported in the original studies. The mean predictions of the CV and test set were evaluated to compare the results of both types of models. In all examples, standard deviations of the predictions were low and very similar for both types of models (Figures S6–S10, bottom).

Importantly, the purpose of this benchmark is to evaluate how models trained with our collection of cost-effective descriptors compare with established DFT-based, publication-quality results. These DFT-based models serve only as benchmarks for the reliability and interpretability commonly accepted in modern molecular ML studies. Therefore, this comparison should not be interpreted as a direct performance competition using identical xTB and DFT descriptors, where higher-level theory would generally be expected to yield better results.

In organometallic reactions A and B, the resulting cost-effective models reduced the three initial DFT descriptors to two xTB descriptors. Despite this reduction, the cost-effective models maintained comparable scaled RMSE values (between 0.9 and 2.6%, Figure 4, bottom left and center) and ROBERT scores (4 vs 5 and 7 vs 7, Figure 4, bottom right). For the cases combining multiple reaction components, C and D, the models reduced the four DFT descriptors originally used to three and two xTB descriptors, respectively. As in the previous cases, xTB- and DFT-based models showed comparable performance in terms of scaled RMSE (from 0.0 to 5.4%) and ROBERT score (5 vs 5 and 9 vs 8). In example E featuring boryl radicals, the reduction in the number of required features was more pronounced, from six DFT to two xTB descriptors. Albeit this substantial reduction, both models performed similarly in terms of scaled RMSE (3.0 and 0.0%) and a slightly better ROBERT score for the xTB-based model (7 vs 5).

We further analyzed differences in error performance between the models using paired *t*-tests[74] and Wilcoxon signed-rank tests (Table S4).[75] In these tests, the RMSE (A–D) or accuracy (E) values from each of the 50 folds generated by the 10× 5-fold CV were compared between the two types of models. The results indicate an overall comparable error performance (Figure 4, bottom left and center), with six out of 10 cases showing no statistically significant differences (gray squares), three cases showing slightly better scaled RMSE values up to 3.0% for the DFT models (red triangles) and one case showing a 5.4% lower RMSE for the cost-effective models (green triangles). The results obtained using mean absolute error (MAE) values led to similar conclusions (Table S5). These findings are consistent with the comparable ROBERT scores obtained: examples B and C show identical values, D favors the DFT models, and A and E favor the cost-effective models (Figure 4, bottom right). Overall, the results indicate a largely equivalent performance scenario between the cost-effective models and their DFT-based analogues.

Importantly, the descriptor generation workflow required only 8–73 min using 8 processors (Figure 4), whereas typical DFT workflows would require days to weeks. For comparison, we performed selected DFT calculations on the same computer architecture and observed reductions of approximately two to four orders of magnitude in computational time when using xTB (Table S3). This drastic reduction is essential for alleviating the often-observed bottlenecks originating from descriptor generation in contemporary QM-based ML modeling.

To assess how the chemistry community perceives these results, we conducted a blinded survey in which experts in ML and catalysis evaluated, side by side, cost-effective and DFT-based models for the five case studies of Figure 4. Two categories, reliability and interpretability, were evaluated on a scale from 1 (poor) to 5 (excellent), followed by a final question assessing overall preference. The survey included a diverse group of respondents, with 52 researchers from 40 institutions across 13 countries (Table S6). We believe that this is a reasonable number and diversity of researchers to form a representative group with varied biases, backgrounds and experience levels, allowing us to assess whether our cost-effective models are likely to be accepted by the broader chemistry community.

The participants were provided with multiple elements to support their evaluations, including reaction schemes, explanations of descriptors, parity plots, model performance metrics from cross-validation and test sets, and SHAP feature analyses. To avoid potential biases that could disadvantage cost-effective approaches, no labels indicating the level of theory used to generate the descriptors were provided. Comprehensive information about the survey, along with the responses and a copy of the survey, is provided in Figures S12–S14 and Tables S6 and S7.

To analyze the results, we first used *t*-tests and Wilcoxon signed-rank tests to assess whether differences in the side-by-side rankings were statistically significant at the 95% confidence level. Both analyses yielded similar conclusions (Table S7). In three cases (A, D, and E), no statistically significant differences in reliability were observed, while in the remaining cases, preferences alternated between DFT-based and cost-effective models (Figure 5, top). For interpretability, researchers rated both sets of models similarly across all five cases, with no statistically significant differences (Figure 5, middle). Likewise, when participants were asked to indicate their preferred model, we observed alternating acceptance of cost-effective and DFT-based models in cases A, C, D, and E and a 50:50 preference ratio in case B (Figure 5, bottom). It is worth noting that the survey included a substantial proportion of experts with more than five years of experience and that the observed trends were consistent across experience levels (Figures S12 and S13).

These results suggest that, despite being generated faster and requiring fewer descriptors, the quality of the proposed cost-effective models is perceived as comparable to that of their published DFT-based counterparts. This outcome can be attributed to the advantages that the proposed cost-effective workflow offers over common strategies used in QM-based chemical ML. First, while descriptor generation becomes several orders of magnitude faster (Table S3), the resulting descriptors often show good correlation with DFT-derived values and capture similar trends (Figure 2B). Another important advantage is that the workflow generates a diverse set of 39 features that are not often considered together, increasing the chances of identifying meaningful descriptors relevant to the target ML problem.

It is worth mentioning that the survey is intended to provide a qualitative assessment demonstrating that, when evaluated blindly without labels indicating the level of theory, the developed cost-effective ML models can compete with DFT-based

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565–12574

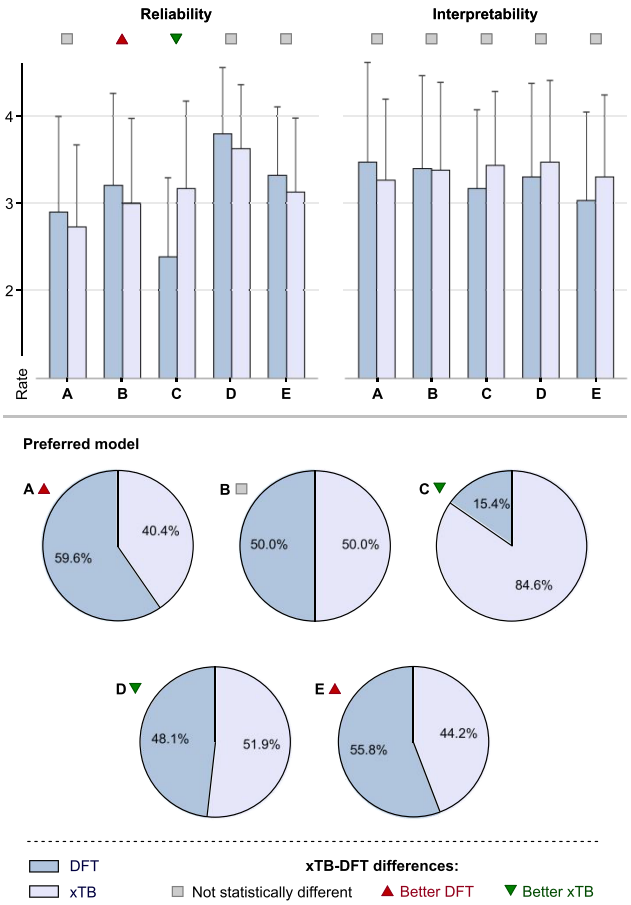


Figure 5. Survey results from 52 experts, including average ratings and standard deviations on a Likert scale from 1 to 5 for reliability and interpretability (top) and overall model preference (bottom). Gray squares denote results that are not statistically different (based on *t*-tests and Wilcoxon signed-rank tests for reliability and interpretability).

counterparts at publication quality. Attempting to quantitatively define an exact "model quality" metric in chemistry or homogeneous catalysis is likely impractical, given the vast diversity of applications, subfields, and researcher backgrounds, and is therefore beyond the scope of this work.

The generality of the proposed descriptors was further assessed by comparing ML models built with our workflow against five additional reported DFT-based models (Figure S11). The benchmark covered a wide range of low- and mid-data regimes, included heavy transition metal catalysts, and encompassed models in areas beyond catalysis. As in the previous examples, the cost-effective and DFT-based models showed similar RMSE values and ROBERT scores, further supporting that the generated cost-effective descriptors are broadly applicable to chemical ML problems. Nevertheless, it is important for researchers to remain aware of the intrinsic limitations of the methods employed and to avoid their use in cases where they are known to be unreliable (Figure 2C).

The similar results and survey conclusions obtained using both types of descriptors suggest that cost-effective features can serve as a feasible and general alternative to DFT in contexts where DFT has historically been preferred. In this regard, we have successfully employed cost-effective descriptors in multiple data-driven studies without incorporating DFT features, supporting their broader adoption. Examples include the active learning-driven discovery of Pd chromophores,[37] reactivity analysis of alkyne semihydrogenation,[76] selectivity trends in Rh catalysts,[77] and activation barrier prediction in epoxide-based cycloaddition reactions.[78]

Fully Automated SMILES-to-Prediction Workflow

To further extend the applicability of the proposed workflow, we provide a fully automated SMILES-to-prediction pipeline designed to generate predictive ML models with minimal user intervention. One of the main challenges limiting the broader adoption of ML in chemical research is the lack of accessible protocols for building models without requiring extensive expertise in dataset construction, descriptor selection, and model screening. In this context, the proposed workflow provides a practical and reproducible strategy for model generation starting from simple molecular inputs.

The workflow combines the descriptor generation protocols presented in Figure 2 with ROBERT tasks, allowing chemists to move directly from a CSV dataset of SMILES strings to ML predictors. The entire SMILES-to-prediction framework is executed automatically through ROBERT using a single command-line instruction or via the easyROB graphical interface, which can also handle ChemDraw file uploads as inputs.[79] The program then enables end-to-end ML modeling without manual intervention, encompassing data curation, hyperparameter optimization, model screening, and other steps typically required in QM-based ML workflows. After model training, the resulting model can be used to predict new outcomes by supplying further CSV files with SMILES strings.

The generality of the SMILES-to-prediction workflow was assessed through three tasks spanning different chemical domains. First, the workflow was applied to the prediction of enantioselectivity in the organocatalytic reaction dataset of case D, comprising 153 data points (Figure 6A).[69] Two separate SMILES columns corresponding to the substrate and the chiral organocatalyst were used as inputs, and descriptors for each component were generated and combined automatically. The procedure then trained a gradient boosting model, which achieved strong predictive performance in both 10× 5-fold CV (R^2 = 0.88, scaled RMSE = 7.7%) and test set (R^2 = 0.93, scaled RMSE = 5.1%) while using only 10 descriptors. Consistently, the ROBERT score showed a high value of 9. These results are comparable to those of the reported DFT-based model (scaled RMSE of 8.7% and 6.1% for CV and test sets, respectively, and ROBERT score = 9, Figure 4D).

The protocol was next applied to predict the aqueous solubility of organic molecules using the ESOL database with 1124 entries (Figure 6B).[80] The automated pipeline generated a neural network model with 9 descriptors that exhibited robust predictive performance in both 10× 5-fold CV (0.85, scaled RMSE = 5.8%) and test set (R^2 = 0.86, scaled RMSE = 5.8%), achieving a ROBERT score of 9.

Lastly, the workflow was applied to the prediction of activation barriers for hydrogenation reactions catalyzed by Vaskas' complexes, using a dataset of 1901 SMILES entries (Figure 6C).[81] In this case, ROBERT identified a neural network model with 20 descriptors as optimal, yielding consistent performance across 10× 5-fold CV (R^2 = 0.87, scaled RMSE = 6.3%) and test set (R^2 = 0.88, scaled RMSE = 6.7%), and a ROBERT score of 8.

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565−12574

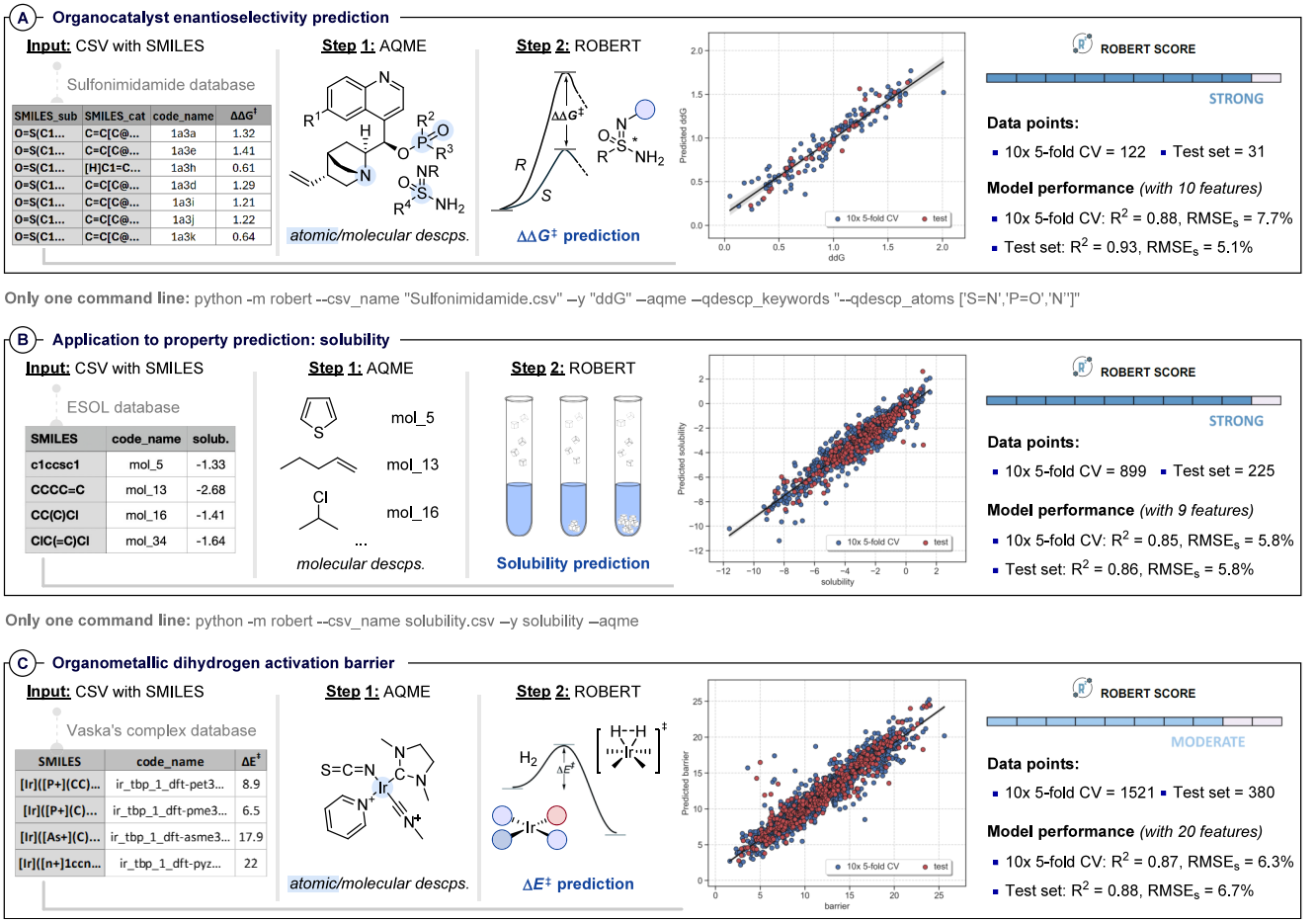


Figure 6. Automated SMILES-to-prediction workflows A–C, including performance metrics of the resulting models and the command-line instruction executed in each case. $RMSE_s$: scaled RMSE. Workflow followed: automated (i) SMILES-to-descriptor generation with AQME and (ii) feature selection and ML modeling with ROBERT.

Across all three case studies, spanning multicomponent stereoselectivity, organometallic reaction barriers, and bulk property prediction, the workflow delivers predictive models with minimal user intervention and execution times compatible with day-to-day research ($A = 0.6$ h, $B = 1.3$ h, and $C = 7.1$ h using 8 processors). Importantly, the same automated protocol is applied across chemically distinct problems, underscoring the generality of low-cost QM descriptors combined with end-to-end ML automation. Together, these examples illustrate how fully automated SMILES-to-prediction pipelines can transform ML from a specialist tool into a practical component of routine chemical discovery.

Despite the advantages and accessibility of this fully automated workflow, users should be aware that it is designed to maximize predictive performance. Although the number of descriptors used for model training is kept low through ROBERT routines,[71] the resulting predictive models may be difficult to interpret since descriptors are automatically selected using recursive feature elimination with cross-validation (RFECV)[67] rather than through manual curation. Consequently, this strategy is most suitable for tasks in which ML is primarily used to guide experimental decision-making in a partially or fully black-box manner. In a similar context, the lack of manual feature selection complicates the application of this workflow to small datasets. In our experience, datasets with fewer than 100 data points often benefit from a manual feature selection step prior to training ML models with ROBERT, whereas larger datasets are much less affected.

■ CONCLUSIONS

In this work, we have introduced an automated and user-friendly AQME workflow for the generation of 39 cost-effective descriptors and their integration into ML models. Relying on xTB and MORFEUS methods, the proposed approach significantly reduces computational cost compared to traditional DFT-based descriptor generation, while maintaining strong correlation with DFT trends across a broad spectrum of organic and organometallic compounds.

Importantly, since ML models rely more on relative trends than on absolute values, and because the workflow introduces a broader variety of descriptors compared to traditional studies, these cost-effective descriptors provide a practical alternative to DFT for data-driven chemical discovery. Beyond technical validation, a blinded survey involving 52 participants reveals strong community acceptance of the approach. Experts judged ML models built from cost-effective descriptors to be comparable in reliability and interpretability to publication-quality DFT-based models. This outcome highlights not only the robustness of the descriptors but also their potential to foster broader

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565–12574

adoption of ML methodologies in homogeneous catalysis, drug discovery and other fields involving molecular systems.

Finally, the introduction of a fully automated SMILES-to-predictor protocol lowers the barrier to entry for ML-driven chemistry. By enabling rapid, systematic, and reproducible model development directly from SMILES strings, this workflow aims to democratize access to QM-based ML predictors and promote the integration of ML into routine chemical research.

METHODS

Molecular descriptors were generated with AQME version 2.0.0, which uses xTB version 6.7.1 and MORFEUS version 0.8.0. Data curation, model screening, hyperoptimization, and all other tasks related to ML modeling were performed with ROBERT version 2.1.0. Installation instructions and details on how to use the programs, along with case studies, are provided in the Supporting Information.

Workflows were executed at CESGA using 8 processors (Intel Xeon Ice Lake 8352Y architecture), unless otherwise specified. All code and raw data used in this study have been deposited in Zenodo to ensure transparency and reproducibility.

ASSOCIATED CONTENT

Data Availability Statement

Raw data and results, instructions, and code used for descriptor generation and ML modeling are freely available on Zenodo (DOI: 10.5281/zenodo.19368369, https://doi.org/10.5281/zenodo.19368369).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.6c02583.

Additional details and methods, including installation instructions, descriptor and workflow details, information on unsupervised and supervised problems, and raw survey data and a copy of the survey (DOCX)

AUTHOR INFORMATION

Corresponding Author

Juan V. Alegre-Requena – *Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC-Universidad de Zaragoza, Zaragoza 50009, Spain*; orcid.org/0000-0002-0769-7168; Email: jv.alegre@csic.es

Authors

David Dalmau – *Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC-Universidad de Zaragoza, Zaragoza 50009, Spain; Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112-0850, United States*; orcid.org/0000-0002-2506-6546

Lauriane Jacot-Descombes – *Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8093, Switzerland; NCCR Catalysis, ETH Zurich, Institute for Chemical and Bioengineering, Leopold-Ruzicka-Weg 4, Zurich 8093, Switzerland*

Adarsh Kalikadien – *Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands*; orcid.org/0000-0002-5414-3424

Brenda Manzanilla – *Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC-Universidad de Zaragoza, Zaragoza 50009, Spain*

Evgeny A. Pidko – *Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands*; orcid.org/0000-0001-9242-9901

Kjell Jorner – *Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8093, Switzerland; NCCR Catalysis, ETH Zurich, Institute for Chemical and Bioengineering, Leopold-Ruzicka-Weg 4, Zurich 8093, Switzerland*; orcid.org/0000-0002-4191-6790

Matthew S. Sigman – *Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112-0850, United States*; orcid.org/0000-0002-5746-8830

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acscatal.6c02583

Author Contributions

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

Funding

J.V.A.-R. and D.D. acknowledge Gobierno de Aragón-Fondo Social Europeo (Research Group E07_23R) and the State Research Agency of Spain (MCIN/AEI/10.13039/501100011033) for financial support (PID2022-140159NA-I00 and AIA2025-163492-C51). D.D. is hired under the Generation D initiative, promoted by Red.es, an organization attached to the Ministry for Digital Transformation and the Civil Service, for the attraction and retention of talent through grants and training contracts, financed by the Recovery, Transformation, and Resilience Plan through the European Union's Next Generation funds (MOMENTUM, MMT24-ISQCH-01).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

J.V.A.-R. and D.D. acknowledge the computing resources at the Galicia Supercomputing Center, CESGA, including access to the FinisTerrae supercomputer and the Drago cluster facility of SGAI-CSIC. The authors also thank Simone Gallarati and Jamie Cadge for helpful discussions related to this work.

REFERENCES

(1) 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.

(2) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **2023**, *14*, 226–244.

(3) 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.

(4) Dalmau, D.; García-Abellán, S.; Alegre-Requena, J. V. Machine Learning in Homogeneous Catalysis: Basic Concepts and Best Practices. *ACS Catal.* **2026**, *16*, 1–11.
(5) Jacot-Descombes, L.; Jorner, K.Predicting Reactivity with Machine Learning. In *Artificial Intelligence in Catalysis*; Ananikov, V. P.; Polynski, M. V., Eds.; Wiley-VCH: Weinheim, Germany, **2025**, pp. 157-194.
(6) 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*, No. 100068.
(7) Abraham, B. M.; Jyothirmai, M. V.; Sinha, P.; Viñes, F.; Singh, J. K.; Illas, F. Catalysis in the digital age: Unlocking the power of data with machine learning. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024**, *14*, No. e1730.
(8) Casillo, E.; Scattolin, T.; Nolan, S. P. Catalysis meets machine learning: a guide to data-driven discovery and design. *Chem. Commun.* **2025**, *61*, 18247–18272.
(9) Kariofillis, S. K.; Jiang, S.; Żurański, 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.
(10) 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.
(11) Tyrin, A. S.; Boiko, D. A.; Kolomoets, N. I.; Ananikov, V. P. Digitization of molecular complexity with machine learning. *Chem. Sci.* **2025**, *16*, 6895–6908.
(12) Deolka, S.; Dalmau, D.; Garcia Roca, A.; Vasylevskyi, S.; Samha, M. H.; Sigman, M. S. Development of a Heteroleptic Palladium Catalyst for the Fluorination of Arylboronic Acids using Data Science Tools. *J. Am. Chem. Soc.* **2025**, *147*, 35705–35717.
(13) Kozlov, K. S.; Boiko, D. A.; Burykina, J. V.; Ilyushenkova, V. V.; Kostyukovich, A. Y.; Patil, E. D.; Ananikov, V. P. Discovering organic reactions with a machine-learning-powered deciphering of tera-scale mass spectrometry data. *Nat. Commun.* **2025**, *16*, 2587.
(14) Burnage, A. L.; Kneiding, H.; Nova, A.; Balcells, D. Cross-coupling catalysts for inactive substrates from a three-dimensional virtual screening. *J. Catal.* **2025**, *452*, No. 116446.
(15) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Martinez Alvarado, J. I.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89–96.
(16) Wu, Y.; Walsh, A.; Ganose, A. M. Race to the bottom: Bayesian optimisation for chemical problems. *Digit. Discov.* **2024**, *3*, 1086–1100.
(17) Ferrer, M.; Deng, B.; Alfonso-Ramos, J. E.; Stuyver, T. Screening Diels−Alder reaction space to identify candidate reactions for self-healing polymer applications. *Digital Discovery* **2025**, *4*, 3400–3410.
(18) Keto, A.; Guo, T.; Underdue, M.; Stuyver, T.; Coley, C. W.; Zhang, X.; Krenske, E. H.; Wiest, O. Data-Efficient, Chemistry-Aware Machine Learning Predictions of Diels−Alder Reaction Outcomes. *J. Am. Chem. Soc.* **2024**, *146*, 16052–16061.
(19) Kurbakov, M. Y.; Sulimova, V. V.; Kopylov, A. V.; Seredin, O. S.; Boiko, D. A.; Galushko, A. S.; Cherepanova, V. A.; Ananikov, V. P. Determining the orderliness of carbon materials with nanoparticle imaging and explainable machine learning. *Nanoscale* **2024**, *16*, 13663–13676.
(20) Mou, L.-H.; Han, T.; Smith, P. E. S.; Sharman, E.; Jiang, J. Machine Learning Descriptors for Data-Driven Catalysis Study. *Adv. Sci.* **2023**, *10*, No. 2301020.
(21) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561–6594.
(22) 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.
(23) Schleinitz, J.; Carretero-Cerdán, A.; Gurajapu, A.; Harnik, Y.; Lee, G.; Pandey, A.; Milo, A.; Reisman, S. E. Designing Target-specific Data Sets for Regioselectivity Predictions on Complex Substrates. *J. Am. Chem. Soc.* **2025**, *147*, 7476–7484.
(24) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.
(25) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; De Bin, R.; Balcells, D. Deep learning metal complex properties with natural quantum graphs. *Digital Discovery* **2023**, *2*, 618–633.
(26) Gee, W.; Doyle, A.; Vargas, S.; Alexandrova, A. N. Multi-level QTAIM-enriched graph neural networks for resolving properties of transition metal complexes. *Digital Discovery* **2025**, *4*, 3378–3388.
(27) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202205735.
(28) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *11*, No. e01493.
(29) Grimme, S.; Bannwarth, C.; Shushkov, P. J. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.
(30) Luchini, G.; Paton, R. S. Bottom-Up Atomistic Descriptions of Top-Down Macroscopic Measurements: Computational Benchmarks for Hammett Electronic Parameters. *ACS Phys. Chem Au.* **2024**, *4*, 259–267.
(31) 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.
(32) Hiener, D. C.; Folmsbee, D. L.; Langkamp, L. A.; Hutchison, G. R. Evaluating fast methods for static polarizabilities on extended conjugated oligomers. *Phys. Chem. Chem. Phys.* **2022**, *24*, 23173–23181.
(33) Gallarati, S.; Bucci, E. M.; Doyle, A. G.; Sigman, M. S. Transferable enantioselectivity models from sparse data. *Nature* **2026**, *651*, 637–646.
(34) Hinostroza Caldas, A.; Kokorin, A.; Tkatchenko, A.; Medrano Sandonas, L. Assessing the performance of quantum-mechanical descriptors in physicochemical and biological property prediction. *Digital Discovery* **2026**, *5*, 803–818.
(35) Jorner, K.; Jacot-Descombes, L.; Turcani , L. kjelljorner/morfeus: v0. 7.2. *Zenodo*.
(36) Alegre-Requena, J. V.; Sowndarya, S.; Pérez-Soto, R.; Alturaifi, T.; Paton, R. S. AQME: Automated Quantum Mechanical Environments for Researchers and Educators. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023**, *13*, No. e1663.
(37) Dalmau, D.; Alegre-Requena, J. V. ROBERT: Bridging the Gap between Machine Learning and Chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024**, *14*, No. e1733.
(38) Dalmau, D.; Alegre-Requena, J. V. Integrating Digital Chemistry within the Broader Chemistry Community. *Trends Chem.* **2024**, *6*, 459–469.
(39) Stefaniu, A.; Pintilie, L. Molecular Descriptors and Properties of Organic Molecules. In *Symmetry (Group Theory) and Mathematical Treatment in Chemistry*. Ed. Akitsu; T. **2018**, chapter 9. IntechOpen: London.
(40) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1874.
(41) Pal, R.; Chattaraj, P. K. Chapter 6 - DFT for reactivity studies/conceptual DFT: History, current state, and perspectives. *Theor. Comput. Chem.* **2025**, *25*, 171–189.
(42) 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.
(43) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9*, 2313–2323.
(44) Dhayalan, V.; Gadekar, S. C.; Alassad, Z.; Milo, A. Unravelling mechanistic features of organocatalysis with in situ modifications at the secondary sphere. *Nat. Chem.* **2019**, *11*, 543–551.

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565−12574

(45) Lau, S. H.; Borden, M. A.; Steiman, T. J.; Wang, L. S.; Parasram, M.; Doyle, A. G. Ni/Photoredox-Catalyzed Enantioselective Cross-Electrophile Coupling of Styrene Oxides with Aryl Iodides. *J. Am. Chem. Soc.* **2021**, *143*, 15873–15881.
(46) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting relative efficiency of amide bond formation using multivariate linear regression. *Proc. Natl. Acad. Sci.* **2022**, *119*, No. e2118451119.
(47) Peñín, B.; Sanosa, N.; Merino-Robledillo, C.; García-Camacho, A.; Sampedro, D.; Hueffel, J. A.; Funes-Ardoiz, I. Mapping Boryl Radical Properties and Reactivity Using Machine Learning: The B-Rad and React-B-Rad Maps. *Angew. Chem., Int. Ed.* **2025**, *64*, No. e202511509.
(48) Landrum, G. *RDKit: Open-Source Cheminformatics*. http://www.rdkit.org (accessed Dec 2025).
(49) For further details, see the AQME documentation: https://aqme.readthedocs.io (accessed Dec 2025).
(50) De Souza, B. GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters. *Angew. Chem., Int. Ed.* **2025**, *64*, No. e202500393.
(51) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST—A Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160*, No. 114110.
(52) Rasmussen, M. H.; Strandgaard, M.; Seumer, J.; Hemmingsen, L. K.; Frei, A.; Balcells, D.; Jensen, J. H. SMILES All around: Structure to SMILES Conversion for Transition Metal Complexes. *J. Cheminform.* **2025**, *17*, 63.
(53) Grimme, S.; Müller, M.; Hansen, A. A Non-Self-Consistent Tight-Binding Electronic Structure Potential in a Polarized Double-ζ Basis Set for All *spd*-Block Elements up to $Z = 86$. *J. Chem. Phys.* **2023**, *158*, No. 124111.
(54) Cadge, J. A.; Hart, S. D.; Walroth, R. C.; Mack, K. A.; Sigman, M. S. Bisphosphine ligand conformer selection to enhance descriptor database representation: improving statistical modelling outcomes. *Chem. Sci.* **2025**, *16*, 20473–20485.
(55) Poddar, A.; Chordia, A.; Chattaraj, P. K. QSPR models for n-octanol/water partition coefficient and enthalpy of vaporization using CDFT and information theory-based descriptors. *J. Chem. Sci.* **2024**, *136*, 23.
(56) Modak, A.; Alegre-Requena, J. V.; De Lescure, L.; Rynders, K. J.; Paton, R. S.; Race, N. Homologation of Electron-Rich Benzyl Bromide Derivatives via Diazo C−C Bond Insertion. *J. Am. Chem. Soc.* **2022**, *144*, 86–92.
(57) Allam, O.; Cho, B. W.; Kim, K. C.; Jang, S. S. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. *RSC Adv.* **2018**, *8*, 39414.
(58) Popescu, M. V.; Parker, N. A.; Jia, Z.; Solon, P.; Maloney, C. J.; Alegre-Requena, J. V.; Paton, R. S.; Smith, M. D. Photocatalysis as a mechanistic probe for the Staudinger β-lactam synthesis. *Chem Catal.* **2025**, *5*, No. 101493.
(59) Froitzheim, T.; Müller, M.; Hansen, A.; Grimme, S.G-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method For the Elements H to Lr (Z=1−103). **2025**.
(60) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29*, 6245–6258.
(61) Gensch, T.; dos Passos Gomes, G.; 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.
(62) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.
(63) McInnes, L.; Healy, J.; Astels, S. HDBSCAN: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205.
(64) Bonakala, S.; Aupetit, M.; Bensmail, H.; El-Mellouhi, F. A human-in-the-loop approach for visual clustering of overlapping materials science data. *Digital Discovery* **2024**, *3*, 502–513.
(65) Blanco-Portals, J.; Peiró, F.; Estradé, S. Strategies for EELS Data Analysis. Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering. *Microsc. Microanal.* **2022**, *28*, 109–122.
(66) Strehl, A.; Ghosh, J. Cluster Ensembles− A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
(67) 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, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(68) Fang, C.; Fantin, M.; Pan, X.; De Fiebre, K.; Coote, M. L.; Matyjaszewski, K.; Liu, P. Mechanistically Guided Predictive Models for Ligand and Initiator Effects in Copper-Catalyzed Atom Transfer Radical Polymerization (Cu-ATRP). *J. Am. Chem. Soc.* **2019**, *141*, 7486–7497.
(69) Haas, B. C.; Lim, N.-K.; Jermaks, J.; Gaster, E.; Guo, M. C.; Malig, T. C.; Werth, J.; Zhang, H.; Toste, F. D.; Gosselin, F.; Miller, S. J.; Sigman, M. S. Enantioselective Sulfonimidamide Acylation via a Cinchona Alkaloid-Catalyzed Desymmetrization: Scope, Data Science, and Mechanistic Investigation. *J. Am. Chem. Soc.* **2024**, *146*, 8536–8546.
(70) Shalit Peleg, H.; Milo, A. Small Data Can Play a Big Role in Chemical Discovery. *Angew. Chem., Int. Ed.* **2023**, *135*, No. e202219070.
(71) 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.
(72) Wu, Z.; Zhu, M.; Kang, Y.; Leung, E. L.-H.; Lei, T.; Shen, C.; Jiang, D.; Wang, Z.; Cao, D.; Hou, T. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Brief. Bioinform.* **2021**, *22*, No. bbaa321.
(73) For more information about ROBERT protocols, see its documentation in Read the Docs: https://robert.readthedocs.io/en/latest/ (accessed Dec 2025).
(74) Student The probable error of a mean. *Biometrika* **1908**, *6*, 1–25.
(75) Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1945**, *1*, 80–83.
(76) Martinez-Fernandez, M.; Yeamin, Md. B.; Dalmau, D.; Carbó, J. J.; Poater, A.; Alegre-Requena, J. V. Data-Driven Analysis of Ni-Catalyzed Semihydrogenations of Alkynes. *Adv. Syn. Catal.* **2025**, *367*, No. e202401444.
(77) Poater, A.; García-Abellán, S.; Alegre-Requena, J. V.; Trzaskowski, B.; Martínez, J. CatalySeed − A Reaction Database for Ruthenium-Catalyzed Ethenolysis of Seed Oils with Applications in Machine Learning. *ACS Catal.* **2026**, *16*, 2160–2170.
(78) Ortiz-García, T.; Posada-Pérez, S.; El-Khchin, L.; Dalmau, D.; Alegre-Requena, J. V.; Solà, M.; D'Elia, V.; Poater, A. Systematic investigation of the role of the epoxides as substrates for CO2 capture in the cycloaddition reaction catalysed by ascorbic acid. *Ind. Chem. Mater.* **2025**, *3*, 452–463.
(79) For a tutorial to install and used the easyROB interface, see the videos available at our YouTube channel: www.youtube.com/@thealegregroup4964 (accessed Dec 2025).
(80) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
(81) Friederich, P.; Passos Gomes, G. dos; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

https://doi.org/10.1021/acscatal.6c02583
ACS Catal. 2026, 16, 12565−12574