Check for updates

Chemistry – A European Journal

ChemEurJ Chemistry Europe European Chemical Societies Publishing

PERSPECTIVE OPEN ACCESS

The Future of Foundation Machine Learning Potentials and DFT in Homogeneous Catalysis: Competition or Synergy?

Maxime Ferrer[1] | Julen Munarriz[2] | Thijs Stuyver[1] | Ruben Laplaza[3]

[1]Ecole Nationale Supérieure de Chimie de Paris, CNRS, i-CLeHS, Paris, France | [2]Departamento de Química Física and Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza, Spain | [3]IIQ, Instituto de Investigaciones Químicas (CSIC-Universidad de Sevilla), Consejo Superior de Investigaciones Científicas, Seville, Spain

Correspondence: Thijs Stuyver (thijs.stuyver@chimieparistech.psl.eu) | Ruben Laplaza (ruben.laplaza@iiq.csic.es)

Received: 26 February 2026 | **Revised:** 30 March 2026 | **Accepted:** 2 April 2026

Keywords: catalysis | computational chemistry | DFT | homogeneous catalysis | ML

ABSTRACT

While DFT is the computational method of choice for mechanistic insight in homogeneous catalysis, the recent rise of foundation-level machine learning interatomic potentials (MLIPs) invites reconsideration: are we approaching competition, or a deeper synergy? These pretrained, fast surrogates are able to map reaction space, sample conformers, and flag likely transition states, potentially displacing routine low-level DFT. Yet their reliability hinges on calibrated uncertainty, transferability across ligand and oxidation-state manifolds, and faithful treatment of long-range polarization, solvation, and open-shell or multireference character. We argue that the near future will likely be contested: MLIPs will handle everyday exploratory tasks, while DFT and higher-level methods will anchor electronic effects, validate high-stakes predictions, and resolve edge cases. If supported by FAIR catalysis datasets, standardized workflows, and robust error quantification, the two approaches will coevolve, enabling scalable, predictive discovery without sacrificing rigor or interpretability.

1 | Introduction

Density functional theory (DFT) remains the workhorse for mechanistic analysis and rational design in homogeneous catalysis [1–3], enabling free-energy profile computations, selectivity rationalizations, and microkinetic models that bridge computation and experiment [4, 5]. In organometallic chemistry, these capabilities often guide hypothesis generation and reaction optimization. However, longstanding limitations still constrain scope and reliability. Results depend on the functional and basis set, finite-temperature and solvation corrections remain fragile, open-shell and multireference regimes remain challenging, and the computational cost rises steeply with system size and conformational complexity [6–12]. These issues become acute when competing spin manifolds or subtle electronic effects control selectivity, which motivates both wavefunction benchmarks for accuracy and approximate semiempirical methods for throughput [13–17].

Foundation-level machine-learning interatomic potentials (foundation MLIPs) add a new axis to this landscape [18–20]. By this we mean models trained on broad, heterogeneous datasets, that aim to capture transferable patterns of bonding and reactivity, in contrast to traditional MLIPs constructed for a specific catalyst or reaction. Trained on large quantum-chemical datasets, modern MLIPs often approach DFT-level energies and forces at much lower computational cost, although performance varies across chemical space and zero-shot transfer remains uneven [21–24]. It is also important to note that semiempirical methods are an active area of research and may offer competitive gains in both speed and accuracy [25–27], sometimes incorporating core machine-learning tools into their parameterization as well [28, 29]. Here,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Chemistry – A European Journal* published by Wiley-VCH GmbH

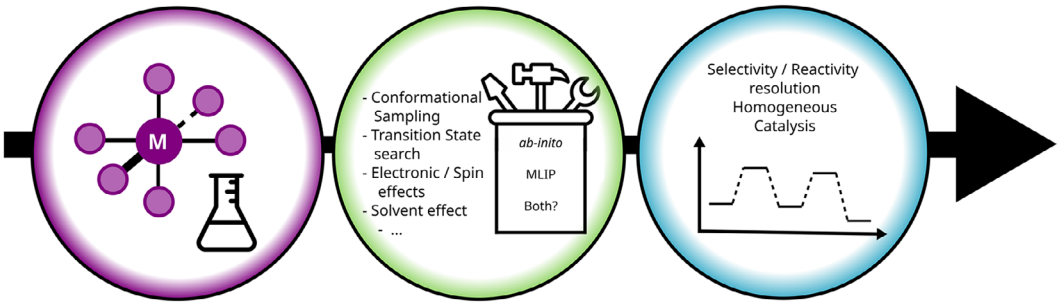

FIGURE 1 | Conceptual landscape of competition and synergy between foundation machine-learning potentials and ab initio methods in homogeneous catalysis.

however, we use the term MLIPs specifically to denote models that do not include any explicit quantum-chemical equations, even in approximate form.

For computational catalysis, the implication is straightforward. If a given foundation MLIP performs reliably within a target domain, the main bottleneck shifts from electronic-structure evaluation to large-scale sampling and exploration. Conformational analysis, reaction-space mapping, and transition-state candidate identification are areas where MLIPs could replace routine DFT steps [30–32]. On the other hand, synergy remains unavoidable wherever electronic structure controls the outcome. Selectivity-determining barriers, spin crossings, charge transfer, and multireference character require diagnostics and controlled accuracy from DFT and higher-level methods. A further complication is that most foundation MLIPs are trained on DFT data, and therefore systematic DFT biases may unpredictably propagate into ML predictions. Reliability depends on calibrated uncertainty and explicit escalation pathways [33–35].

Whether this emerging division of labor leads to the displacement of routine DFT or to co-evolution with electronic-structure methods will depend on usability, data quality, physics coverage, and software infrastructure. Catalysis-relevant training and benchmark sets must include transition-metal chemistry with explicit charge and spin labeling, as well as reactive, off-equilibrium configurations that extend beyond current FAIR efforts. In parallel, model developments that better capture long-range electrostatics, polarization, and solvation are essential for realistic solution-phase catalysis [36, 37]. Finally, broad adoption will require interfaces and standardized workflows that integrate uncertainty, escalation, and reproducibility, analogous to the software ecosystems that underpin modern quantum chemistry [38, 39].

In this perspective, we take a speculative, forward-looking stance. Our goal is not to declare a winner between MLIPs and DFT, but to outline the conditions under which competition becomes credible and the circumstances in which tight coupling is indispensable. The sections that follow examine how reliability should be quantified and managed, the present limitations and bottlenecks, how catalysis-specific datasets and benchmarks must evolve, and how model outputs may interface with microkinetic modeling and experiment to yield trustworthy, actionable predictions (Figure 1).

2 | The Arrival of Foundation-Level ML Potentials: Reality Check

For many catalysis practitioners, "machine-learning potentials" refers to bespoke models trained for a single system, such as one catalyst, one solvent, and one reaction, with substantial data generation and expert effort [36, 41–44]. Foundation MLIPs change this picture [18–20, 23]. Rather than training end to end for a narrow target, a foundation MLIP is pretrained on an extremely large and heterogeneous set of quantum-mechanical reference calculations. It learns reusable representations of chemical environments and their energies and forces. The practical promise is zero-shot use, or light adaptation, on new systems by leveraging patterns learned across molecules, materials, and catalytic motifs including bond formation and breaking.

A useful analogy is a pretrained large language model. Pretraining does not guarantee perfect performance on every new topic, yet it provides a strong prior that transfers across tasks. In atomistic modeling, that prior is an efficient mapping from structure to energies and forces that supports geometry optimization, molecular dynamics, and large-scale sampling at costs far below routine DFT. What distinguishes foundation MLIPs from earlier generations is not the idea of learning a potential energy surface, but the scale and diversity of training data and the resulting expectation of broader transferability (Figure 2) [40, 45, 46].

Recent releases make the foundation idea tangible. The Open Molecules 2025 (OMol25) database provides consistent, hybrid-DFT-quality energies and forces for more than 10^8 systems, and it stores total charge and spin multiplicity [47, 48]. OMol25 is both large and diverse, and it includes substantial non-equilibrium coverage rather than only optimized minima [47, 48]. This matters because potentials used in dynamics must remain stable not only near equilibrium, but also for distorted configurations and reactive regions that trajectories explore.

Building on multi-domain data, Meta's Universal Models for Atoms (UMA) exemplify a foundation MLIP trained at unprecedented scale across molecules, materials, and catalysis [23]. The aim is to amortize data generation and model development by training on hundreds of millions of distinct 3D structures from multiple domains, so that one model supports tasks that previously relied on specialized MLIPs [23]. A foundation MLIP

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

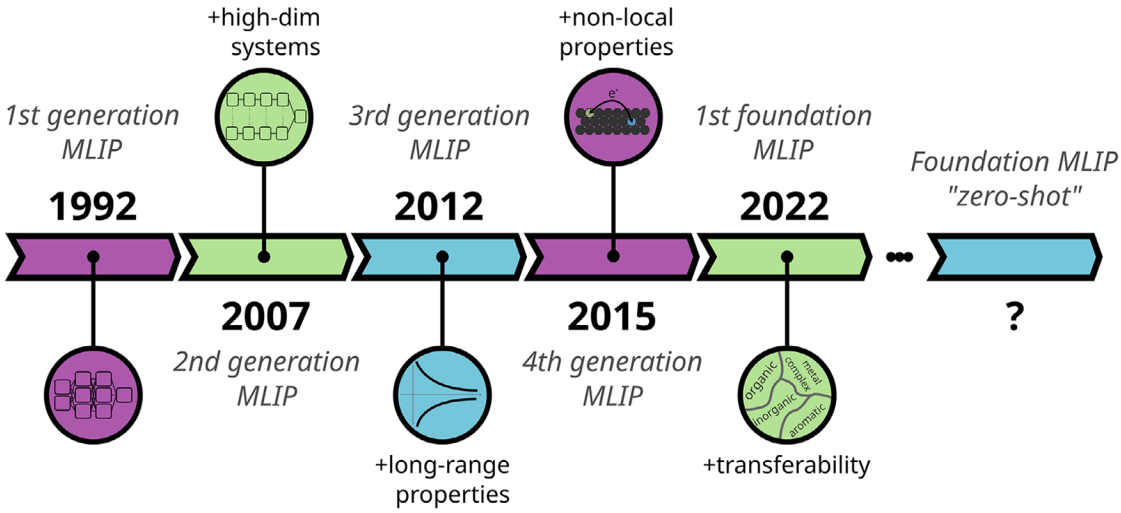

FIGURE 2 | Chronological evolution of machine-learning interatomic potentials, inspired by the MLIP classification proposed by Behler [40].

is not a chemistry oracle, but a broadly pretrained surrogate for DFT energies and forces that aims to be generally usable.

This reality check has two parts. First, "DFT-level" claims for foundation MLIPs usually refer to energies and forces on benchmark test sets. This often suffices for high-throughput conformer sampling, reaction-space exploration, and transition-state candidate generation, yet it does not imply chemically accurate barriers, selectivities, or robust behavior under large distribution shifts, such as modeling novel chemical reactions [18, 23].

Second, transfer is conditional. Homogeneous catalysis spans changes in oxidation state, spin, coordination topology, and solvation environment, so zero-shot reliability depends on whether these motifs appear in pretraining and whether the model captures the physics needed for smooth extrapolation [23, 47]. This motivates a guiding principle used throughout this Perspective: foundation MLIPs are best treated as reusable priors that replace routine exploratory steps in well-known chemical domains, while DFT and higher-level methods remain essential when electronic structure is decisive or when uncertainty indicates extrapolation. Therefore, the challenge for practitioners in the near future is judiciously navigating between these two approaches.

3 | Where Competition is Credible

The strongest case for competition between MLIPs and DFT in homogeneous catalysis is the replacement of routine scouting, where cost is dominated by sampling rather than single-point fidelity. Foundation MLIPs raise the throughput of large-scale conformer enumeration, reaction-space mapping, and transition-state proposal generation to levels that remain prohibitive for modest DFT settings [23, 47, 48]. Additionally, MLIPs running on GPUs benefit from large-scale simulations (in terms of atoms) or trivially parallel tasks, where the full memory of the chip is used during runtime. In this breadth-first regime, the key question is not universal accuracy, but rather whether the model ranks and prioritizes a small set of chemically decisive structures well enough for later DFT adjudication.

Credibility requires an explicit applicability domain defined by elements, key ligand motifs, and relevant charge and spin states. In practice, the first step is domain declaration rather than geometry optimization. The declaration states which metals and ligands, charge and spin states, solvent motifs (if any), and which reference DFT level the workflow aims to emulate. This mirrors standard practice in DFT, where method choice depends on the problem class [2, 6], albeit with additional importance.

Within a declared domain, three task types support ML-first replacement of routine DFT or semiempirical methods. Conformer and speciation sampling benefit because flexible ligands, coordination isomers, ion-pairing motifs, and weak encounter complexes demand ensemble coverage. When the goal is low-lying structures and relative populations, rather than sub-kcal mol^{-1} barrier differences, MLIPs replace most scouting optimizations and enable sampling workflows that would otherwise be truncated [30–32, 49]. Reaction–space mapping and pathway triage benefit because screening plausible routes, such as alternative insertions, ligand exchange, and variants of β-H elimination, often requires dozens of optimizations. Transition-state optimization requires locating saddle-like geometries and reaction coordinates in the potential energy surface, tasks that are typically hampered by the high cost of computing first and second derivatives of the energy with DFT. Owing to their differentiability, MLIPs provide hessians at a low cost [50, 51], which open the door to thorough saddle point optimization algorithms and quick, effective exploration [52].

Operationally, MLIP outputs are best treated as high-coverage proposals with safeguards rather than direct replacements for DFT. Trust is highest within the declared domain and after basic stability checks during short optimizations or short dynamics. Trust improves when these runs show no force spikes, no systematic energy drift, no unphysical rearrangements, and no persistent nonconvergence. The choice must match the task. Substitution is most defensible for ranking, sampling, and proposal generation, not for mechanistic claims that hinge on small energy gaps. A calibration and escalation plan is equally important: uncertainty should be monitored, even through practical proxies, and flagged

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

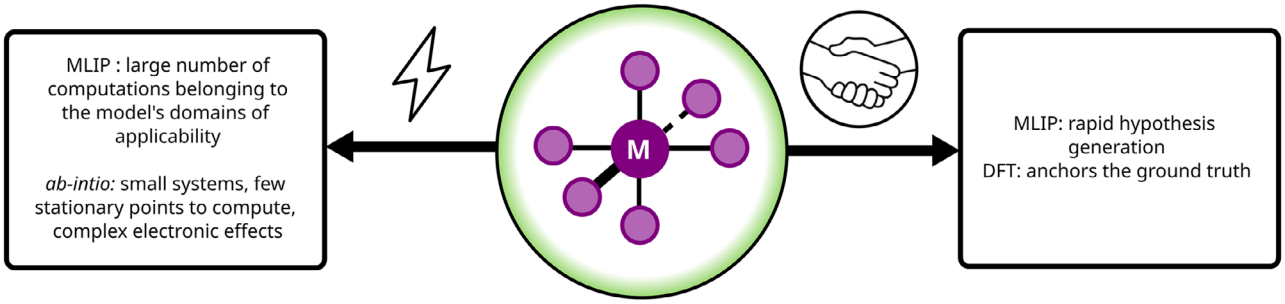

FIGURE 3 | Domains where ML potentials can credibly replace routine ab-initio calculations and where a synergy is inevitable.

structures should move to DFT automatically rather than through ad hoc judgment [34, 35].

Importantly, benchmarking should emulate the workflow at hand, not a static test set. A claim of replacing a routine DFT protocol is falsifiable if the MLIP reproduces the qualitative behavior of that protocol for the structures and distortions produced in use. Validation benefits from task-specific designs, such as conformer ranking, coordination rearrangements, and stability near transition structures, with splits and perturbations that mimic catalytic workflows rather than random holdouts [53]. It must be acknowledged that benchmark construction is not neutral. Curated sets often overrepresent stable, easy structures and understate real-world error. Stress tests that target difficult regions, for example, configurations where methods disagree most, provide a more meaningful safety check for both DFT and MLIP deployment [54, 55].

The practical takeaway is that competition is most credible when MLIPs replace the expensive, repetitive parts of a catalysis workflow, namely scouting, sampling, and proposal generation, while DFT and higher level methods are reserved for electronically decisive points. The goal is fewer DFT calculations per mechanistic conclusion, without lowering standards for validation and interpretability.

4 | Where Synergy is Inevitable

Competition is less compelling when a catalytic conclusion depends on small energy differences driven by electronic effects rather than broad geometric trends. In homogeneous catalysis, selectivity and turnover often depend on barrier differences of only a few kcal mol^{-1}. Foundation MLIPs primarily learn a mapping from atomic structure, and sometimes global labels such as charge and spin, to energies and forces. They therefore rarely provide the electronic diagnostics used to judge whether an energy difference is meaningful. Electronically delicate steps are thus where ML speed should be paired with electronic-structure adjudication rather than used as a drop-in replacement (Figure 3).

This does not make MLIPs irrelevant in high-stakes regions. When the target chemistry lies well within the pretraining manifold, such as common coordination motifs in closed-shell complexes with familiar ligands and charge states, modern MLIPs often reproduce structures and relative energetics well enough to guide exploration and generate hypotheses. However, when mechanisms involve unusual oxidation states, low coordination numbers, pronounced multireference character, explicit charge transfer, or competing spin manifolds, the model prior is often insufficient, and errors become systematic. In these cases, even a plausible ML trajectory may be misleading if the electronic state is incorrect. For mechanistic claims, ML-driven exploration should therefore be coupled to escalation to DFT and higher-level methods. A key difference is that DFT provides diagnostics rooted in physics (e.g., the existence of near-degenerate states or orbitals) that suggest a need for escalation to higher-level methods to knowledgeable practitioners. An analogous diagnosis is not trivial for MLIPs. We return to the hardest electronic regimes and DFT limitations in Section 7.

Synergy is also pragmatic. As systems get smaller and the number of decisive stationary points shrinks, the raw speed advantage of MLIPs matters less. For many homogeneous catalysts with tens of atoms, a modest set of DFT calculations on selectivity-determining steps is already feasible. In that setting, replacing final adjudication with ML adds risk without proportional benefit. The cost landscape is also evolving because GPU-optimized electronic-structure implementations continue to reduce DFT wall times, narrowing the gap for the few computations that decide a mechanism [56, 57].

The most robust near-term strategy is a clear division of labor. MLIPs compete where the goal is breadth. They generate ensembles, map reaction space, and surface candidate structures quickly and cheaply. DFT anchors the points where electronic structure decides the answer and connects with higher-level methods if required. In practice, ML outputs should be treated as high-coverage proposals, while electronic-structure methods are reserved for verification and refinement whenever conclusions require fine energetic discrimination or correct state identity. With explicit escalation criteria, this hybrid approach preserves ML throughput gains without sacrificing the diagnostic richness and mechanistic rigor that homogeneous catalysis demands.

5 | Reliability, Uncertainty, and Escalation

In homogeneous catalysis, speed matters only when predictions are trustworthy. For foundation MLIPs, trust is not captured by a single global MAE or RMSE on random train and test splits because real workflows generate related configurations rather than independent ones. This is most obvious near bond rearrangements, unusual coordination changes, and strained

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

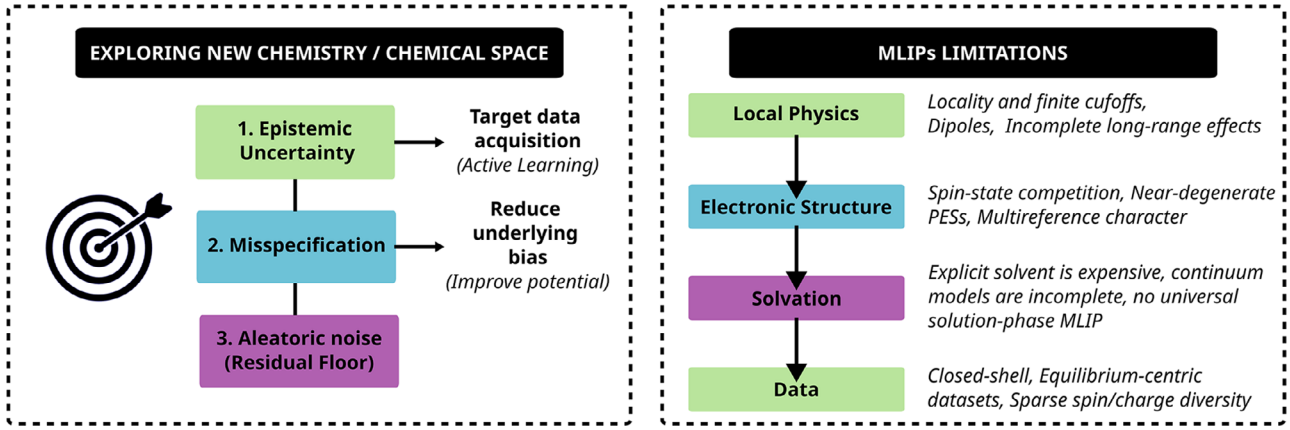

FIGURE 4 | Reliability, uncertainty quantification, and limitations in ML-driven workflows applied to homogeneous catalysis.

geometries. In practice, a workflow needs a warning signal for each structure that indicates when a predicted energy or force is likely to be wrong. This is the role of uncertainty quantification (UQ). The goal is a per-structure estimate of how reliable the predicted energies and forces are, together with evidence that the stated confidence matches the observed error for the deployment domain, including metals, ligands, charge and spin states, and solvation motifs [34, 35].

It helps to separate three sources of uncertainty that behave differently in catalytic workflows (Figure 4). *Aleatoric* uncertainty reflects noise or variability in the reference labels. For foundation MLIPs, the labels are typically DFT energies and forces, so this uncertainty includes the intrinsic scatter and approximations of the chosen DFT protocol. It sets an error floor that no model trained on those labels will systematically beat, and it is best addressed during training with losses that allow structure-dependent noise [34]. *Epistemic* uncertainty reflects lack of coverage. It rises when the model sees chemistries or geometries that differ from what it was trained on, such as new ligand classes, uncommon coordination motifs, changes in oxidation or spin state, rare conformations, or distorted structures near reactive events. Epistemic uncertainty decreases when informative new reference calculations are added. In deployed models, it is often estimated by checking how sensitive the prediction is to small changes in the model, for example, by comparing several independently trained models or by using dropout at prediction time [34, 35]. *Misspecification* arises when the model architecture lacks important physics. In that case, a model may give confident-looking predictions that are systematically biased, for example, when long-range electrostatics and polarization, solvation effects, or spin and charge phenomena are represented poorly [58–60]. In catalysis, epistemic uncertainty often dominates early exploration, while misspecification becomes limiting as coverage increases since better uncertainty estimates do not fix missing physics [58, 59].

While capturing epistemic uncertainty is already an established part of MLIP workflows, it is typically achieved through model ensembles [42]. This has important drawbacks in terms of cost, since several models must be trained and used to predict at each step. Alternatively, there, approaches such as evidential learning, where a single network predicts not only an energy or force but also a distribution that represents confidence, are attracting interest [61]. In molecular property prediction, evidential methods often produce uncertainties that track error and guide data selection at essentially no added prediction cost, which suggests a promising direction for atomistic models when running many models is too expensive.

Because catalytic conclusions often hinge on small energy differences, UQ must be meaningful at the level of decisions. The relevant target is not only uncertainty in an absolute energy, but also uncertainty in barrier differences, free-energy spans, and ensemble-averaged free energies used in microkinetic models [62, 63]. This makes calibration essential. An uncertainty estimate is useful only if, on average, it matches real errors. Raw uncertainties from models are rarely calibrated out of the box, especially under distribution shift, so deployment should include calibration on a small, domain-matched reference set that reflects the configurations the workflow will generate [34]. Practically, one defines the deployment domain and the reference level of theory, builds a calibration set that mirrors typical structures including relevant distortions and solvent or ion motifs, and rescales the raw uncertainty so that stated confidence matches observed error. Post-hoc scaling and conformal approaches are attractive because they are simple and robust, and they provide reliable coverage under mild shifts, particularly when conditioned on whether a structure looks familiar in the model's internal representation [34, 64]. Calibration should be checked both in-domain and on targeted stress tests that mimic catalytic failure modes [65].

Even with calibration, failures are inevitable, so the objective is early detection and selective escalation rather than unconditional trust. In an MLIP-driven workflow, monitoring can be embedded into optimization, sampling, and pathway exploration. ML steps are accepted while uncertainty stays within a predefined envelope, and escalation is triggered when uncertainty rises sharply or when stability diagnostics indicate extrapolation. Practical indicators include unusually large predicted forces, sudden energy changes along short steps, repeated optimization failures, persistent nonconvergence, or geometries that are chemically implausible [65–67]. Flagged configurations are then refined with electronic-structure calculations, typically DFT and sometimes higher-level methods for electronically delicate cases, and the workflow continues using validated results. This escalation loop

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is central to safe use because it makes ML speed conditional on reliability and routes difficult cases to methods that provide electronic diagnostics and controlled accuracy [53, 67–69].

A tension remains between what is ideal for model improvement and what is convenient for practitioners. In bespoke MLIP development, escalation points become new training data in active-learning cycles. For off-the-shelf foundation MLIPs, fine-tuning is often impractical for nonexperts, so escalation may simply mean switching to DFT for problematic segments of a workflow. Even so, flagged structures and validated reference calculations remain valuable. If captured with sufficient metadata and shared in reusable form, they become targeted additions to the next iteration of foundation training corpora and benchmarks, turning deployment failures into systematic coverage expansion (Section 8) [70, 71].

Finally, multi-fidelity strategies offer a complementary route to reliability by separating broad coverage from high-stakes accuracy. Rather than training one potential that is uniformly high fidelity, a practical approach uses abundant DFT-level data for breadth and adds sparse, carefully selected higher-level corrections via Δ-learning on chemically decisive subsets [72]. In catalysis, these subsets include transition-state neighborhoods, spin-state splittings, redox-active motifs, and other regions where small errors have outsized mechanistic impact. The open question is not whether multi-fidelity helps, but how best to integrate it so escalation, calibration, and fidelity selection become routine and standardized rather than ad hoc.

6 | Closing the Gaps in ML Potentials

Foundation-level MLIPs trained across molecules, materials, and catalysis inherit a central limitation of many modern architectures: they are largely local models with finite cutoffs, so performance drops when nonlocal physics controls energies and forces (Figure 4). In homogeneous catalysis this is common, since solution-phase reactivity and charged or open-shell organometallics are often governed by long-range electrostatics, environment-dependent polarization, and non-local charge transfer. Purely local representations struggle to reproduce dielectric screening, field response, and electron redistribution across extended ligand frameworks. Errors may look modest near equilibrium yet become consequential along distorted reaction coordinates, where dipoles, solvation energetics, and barrier heights depend on physics outside a short-range neighborhood [59]. While in some condensed phases long-range information may be transferred through geometric effects and successfully captured by the MLIPs [73, 74], ultimately these are misspecification problems due to local architectures. Even perfect coverage of local geometries does not guarantee a correct response to long-range perturbations [23, 24].

A practical remedy is to augment a short-range MLIP with a self-consistent charge-equilibration layer, such as QEq or kQEq. These models predict environment-dependent electronegativities and solve for a global charge distribution at each step, yielding long-range electrostatics and an implicit form of polarization that improves dipoles, forces, and transferability for ionic and polar systems [75, 76]. However, naive charge equilibration may overpolarize or induce unphysical charge transfer, especially in heterogeneous environments, so successful use typically requires regularization and physically motivated constraints [77]. The broader message is that long-range behavior is rarely learned by scaling local data alone.

Complementary strategies add explicit long-range terms while leaving the MLIP to model chemically complex short-range interactions. One pattern constructs Coulomb interactions from analytic charge distributions, for example associated with ions and Wannier centers, while a neural potential models the short-range remainder, which restores correct tails and improves transfer to larger systems [78]. Another approach treats London dispersion explicitly by learning atomic dispersion coefficients and evaluating the asymptotic term, which improves intermolecular interactions and mitigates a known weakness of purely local regression [79]. These add-ons are attractive because they are orthogonal to dataset scale, and enforcing correct physical limits reduces the burden on the network and improves extrapolation. For a recent perspective on long-range electrostatics for MLIPs, we refer the reader elsewhere [80].

Architectural routes to nonlocality are advancing in parallel. Long-range-aware message passing, attention-based designs, and hybrids with analytic corrections propagate information across a structure more effectively than fixed-cutoff schemes. Some architectures also incorporate charge and spin conditioning to improve state awareness [46, 81]. The likely near-term outcome is not a single winner but robust stacks that combine a strong short-range backbone with explicit electrostatics and polarization and long-range-aware components, tuned to dominant failure modes in the target domain [23, 24, 59, 82].

Solvation compounds these challenges because realistic solution-phase catalysis requires accurate energetics and extensive sampling. Continuum models often miss specific hydrogen bonding, entropic pre-organization, and solvent-mediated charge rearrangements, while explicit solvent introduces large configurational spaces and rare-event sampling barriers. Here MLIPs offer a real opportunity. Low-cost forces make explicit-solvent sampling, enhanced sampling, and larger ensembles practical. Recent demonstrations show that, with active learning, explicit-solvent ML potentials reproduce liquid structure and deliver adsorption and free-energy quantities with near-DFT fidelity at much lower cost [42, 44, 83]. Nonetheless, condensed-phase catalysis remains difficult because pathway exploration is combinatorial, sampling is expensive, and charged spectators such as counterions interact through long-range physics that local models treat poorly [43, 84–86].

Implicit solvation coupled to MLIPs is less mature. A general strategy would mirror DFT practice by combining a transferable gas-phase MLIP with an implicit solvent model. This adds requirements. The MLIP must provide reliable charges and multipoles, or an equivalent electrostatic representation, so the solvation model responds correctly across conformations and charge states. ML-augmented continuum approaches, such as ML-corrected PCM, show that residual learning improves solvation free energies when explicit solvent is infeasible [87], but broadly usable, foundation-level solutions that make off-the-shelf MLIPs reliable for solution-phase catalysis are not yet available.

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

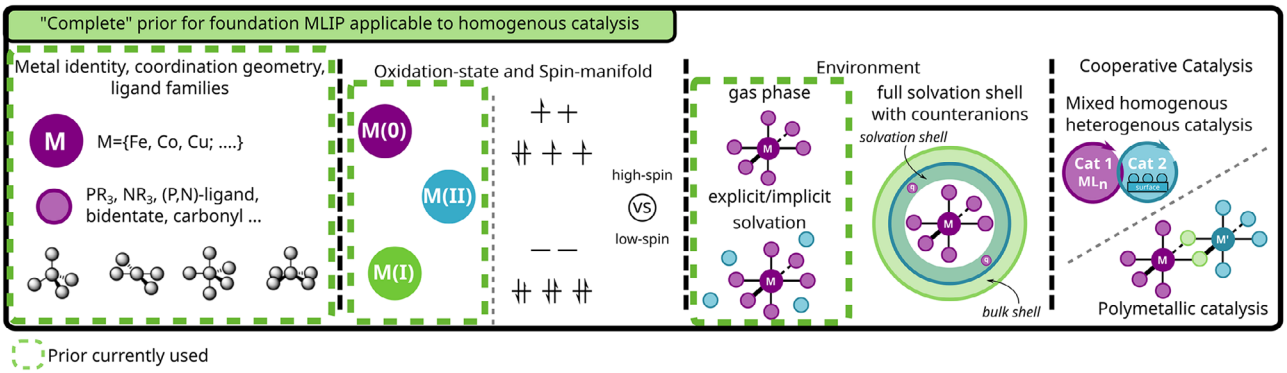

FIGURE 5 | Schematic representation of what would constitute a "complete" prior for the development of a foundation MLIP applicable to homogeneous catalysts, and the current state.

Until this gap closes, solvation remains a bottleneck for deploying foundation MLIPs as default engines for homogeneous catalysis rather than fast scouts [88, 89].

In summary, credible competition with DFT for solution-phase catalysis, charged complexes, and open-shell intermediates requires explicit treatment of long-range electrostatics and polarization and a practical solvation strategy (Figure 5). The most robust near-term stacks will combine a strong short-range MLIP with self-consistent electrostatics, such as charge equilibration, explicit long-range corrections for electrostatics and dispersion, and long-range-aware architectures, validated under the sampling conditions that catalysis demands [23, 24, 59].

7 | The Hard Boundaries

Homogeneous catalysis often operates in an electronic regime where several configurations lie close in energy. First-row transition metals, open-shell intermediates, variable ligand fields, and bond-making and bond-breaking transition states naturally generate near-degeneracies and competing spin manifolds. In these systems, the mechanistically decisive object is often not a single potential energy surface (PES), but the relative placement and coupling of several PESs. Errors of only a few kcal mol^{-1} may reorder spin ladders, shift crossing points, and invert predicted selectivity [13]. This sets a boundary condition for any surrogate that maps structure to energy. When the identity of the relevant electronic state changes along the reaction coordinate, the energy is no longer a single-valued function of geometry unless the state is specified and tracked.

For foundation MLIPs, this issue is acute because most models predict energies and forces from atomic positions, and sometimes global labels such as total charge and spin, without explicit local electronic degrees of freedom. Early generations were effectively insensitive to electron count and spin, which limited use whenever multiple charge or spin surfaces matter, including in redox chemistry [40]. More recent architectures and datasets incorporate total charge and spin information, which is necessary, but it does not resolve the deeper ambiguity that arises when multiple electronic configurations with the same total charge and spin coexist at similar geometries and differ mainly by orbital occupations. In such cases, the mapping from structure to energy becomes effectively multivalued unless the model is conditioned on a richer state description or trained explicitly on multi-state data. Spin-crossover systems and other near-degenerate manifolds, as encountered in many iron and cobalt catalysts with competing high-spin and low-spin pathways, therefore remain difficult for current MLIPs to treat reliably. Related issues appear in redox-active ligands and ligand-to-metal charge-transfer motifs, where small structural changes drive large redistributions of electron density and polarization that are not straightforwardly encoded by local geometric descriptors. Mechanistic conclusions in these regimes depend on electronic diagnostics such as spin densities, charge localization, and orbital occupations, which are native to electronic-structure methods but largely absent from standard MLIP outputs.

This is also where DFT itself becomes a moving target rather than a fixed reference. Spin-state energetics are strongly method-dependent, and credible, broadly applicable reference data are scarce, so reasonable DFT protocols may disagree qualitatively for the same complex [13, 90]. The SSE17 benchmark set, derived from experimental spin-state splittings, illustrates the magnitude of the challenge. Common hybrid and meta-hybrid functionals often show mean absolute errors of about 5–7 kcal mol^{-1} and maximum errors above 10 kcal mol^{-1}. Double hybrids reduce errors, and CCSD(T) approaches chemical accuracy on that set [13]. Even higher-level approaches may remain inconsistent for open-shell transition-metal energetics unless protocols are tightly controlled. The limitation is not simply that DFT fails, but that the electronic structure is intrinsically hard [11–13, 90, 91]. When an MLIP is trained on DFT, it may inherit systematic bias in the reference. When DFT itself is uncertain, ML will not surpass it without targeted higher-level data and careful multi-fidelity design.

Wavefunction methods provide the principled route forward because they represent strong correlation beyond DFT's mean-field framework, but routine use in catalytic workflows remains difficult due to scaling, convergence sensitivity, and the practical reality that multireference and coupled-cluster energetics are often evaluated as single points on DFT geometries rather than through full geometry optimization [90]. A similar boundary appears in photocatalysis and other excited-state mechanisms. The relevant landscape involves multiple electronic states and nonadiabatic couplings, and even within electronic-structure

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

theory the balance between accuracy, robustness, and cost becomes delicate. For MLIPs, the implication is direct. Unless models are trained on multi-state data and equipped to represent state identity and couplings, they should not be expected to reproduce crossings, reordering, or state-specific barrier heights in a reliable and transferable way [92].

A realistic near-term role for ML in these boundary regimes is assisted exploration paired with electronic-structure diagnostics rather than autonomous prediction. Practically, this means using MLIPs to accelerate candidate geometry generation and sampling where they remain stable, while prioritizing escalation to DFT and higher-level methods at points where state competition is expected or where small energetic differences control conclusions. Recent analyses suggest that differences in multireference character between states or structures, often termed multireference imbalance, rather than absolute multireference character alone, are often more predictive of property errors. This provides a principled criterion for where to invest higher-level calculations within multi-level workflows [93]. In the longer term, progress toward foundation MLIPs that remain reliable across strongly correlated regimes will require larger datasets and procedural advances that make higher-level data generation more robust and scalable. Even aside from computational cost, the human and algorithmic burden of generating converged multireference labels for diverse, out-of-equilibrium structures is far higher than for DFT. Progress will likely rely on improved automation, better diagnostics, and multi-fidelity strategies that introduce sparse high-level corrections only where they change decisions [94–97].

In short, the hard boundaries for foundation MLIPs in homogeneous catalysis are defined by state competition and strong correlation. These are situations where the reaction coordinate traverses multiple electronic surfaces, where spin and charge localization change abruptly, or where mechanistic conclusions rely on electronic diagnostics rather than geometry alone. These regimes do not eliminate the value of MLIPs, but they enforce synergy by construction. ML accelerates exploration, while electronic-structure methods remain essential to identify the correct state, quantify decisive energy differences, and provide the diagnostics needed for mechanistic credibility.

8 | Data and FAIR Standards

In spite of the previous sections, the prospects of foundation MLIPs in homogeneous catalysis probably depend less on model class than on the availability of reference data. Key needs are chemically diverse transition-metal and ligand environments, off-equilibrium geometries with forces, and explicit total charge and spin multiplicity [46, 98]. Unlike biomolecular MLIP domains, homogeneous catalysis routinely spans multiple oxidation states, spin manifolds, and coordination topologies, which increases the demands on coverage and electronic-state labeling [46, 82, 99–103].

Because modeling must span several potential energy surfaces across the periodic table, chemical space will not be covered by enumeration, which reaches extraordinary scales even for small organic molecules alone [104]. For MLIPs, useful coverage is instead the diversity of local atomic environments and electronic states visited by simulations. This motivates catalysis benchmarks organized around motifs that matter for dynamics and reactivity, such as coordination changes, bond-making and bond-breaking neighborhoods, and charge and spin variation, rather than static collections of optimized structures alone [105, 106]. Recent benchmarking discussions likewise argue that dataset design and sampling of relevant dynamical modes often dominate downstream simulation reliability beyond what energy and force errors alone suggest [53, 107]. Because molecular MLIP datasets often contain strongly correlated configurations, such as trajectory frames or small distortions, random splits may substantially overestimate generalization.

Transition-metal complexes still lag behind organic molecules in data availability, and large corpora remain essential anchors for representation learning and baseline evaluation. The pioneering `tmQM` dataset provides tens of thousands of mononuclear transition-metal complexes mined from the Cambridge Structural Database (CSD) and labeled with computed properties at standardized levels of theory [108]. Recomputations and curated variants such as `tmQM_wB97MV` address inconsistencies and improve usability [109]. Label-enriched extensions push beyond geometry-only learning. `tmQMg` supplies NBO-informed natural quantum graphs (NatQG) for a large subset, enabling models to exploit chemically meaningful electronic descriptors [110], and `tmQM+` augments `tmQM` with descriptors at multiple levels of theory to probe robustness and transfer to unseen regimes [111]. Application-linked curation, exemplified by `tmCAT`, increases catalytic relevance by identifying catalysis-associated subsets from broad transition-metal databases [112]. Crystallography-mining tools such as `cell2mol` recover connectivity and total charge, including oxidation state, from experimental structures, which helps bridge the CSD to usable datasets [113, 114]. Complementary developments that infer ground-state spin directly from structure, such as TM-GSspin, address another frequently missing label for charge- and spin-aware modeling [115].

For catalysis-facing applications, labels beyond energies and forces are valuable as state descriptors and diagnostics. Consistent total charge and spin, atomic charge and spin populations, dipoles and multipoles, bond-order proxies, or density-derived descriptors help define what is being modeled and enable like-for-like comparisons across datasets and levels of theory [99, 108, 110, 111, 116]. This is already visible in molecular datasets that intentionally expose richer label spaces [99, 108].

For MLIPs, the central gap is also nonequilibrium coverage, meaning forces on distorted geometries resembling what trajectories explore. The original ANI-1 dataset illustrated the importance of large-scale off-equilibrium sampling for molecular MLIPs [117], and in organic chemistry the `Transition1x` database shows how pathway-adjacent sampling can be systematized at scale using NEB-generated reaction paths [106]. Active-learning workflows likewise show that model-guided selection broadens coverage with fewer redundant points than naive sampling, as in `ANI-1x` [105, 118]. Benchmarks should include activated or rare-event configurations, such as transition-state neighborhoods and strained intermediates, since these can control kinetics yet remain poorly learned even when average force and energy errors are low [53, 106]. In practice, stable long-time trajectories may fail despite low test-set errors, often reflecting excursions into poorly

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

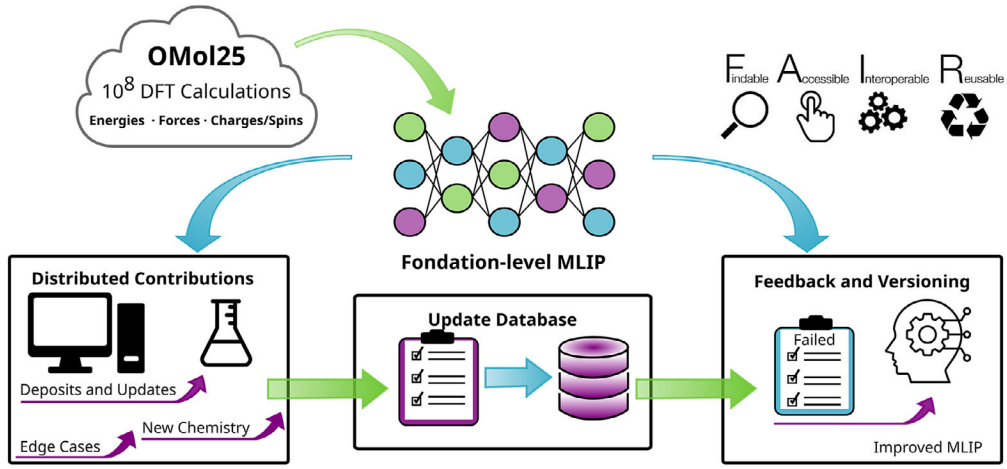

FIGURE 6 | Schematic representation of a potential future for FAIR data infrastructures, illustrated using OMol as an example.

covered regions of configuration space [119]. For homogeneous catalysis, analogous benchmarks should prioritize strained coordination geometries and reaction-center distortions characteristic of organometallic mechanisms, since these regions are sparsely sampled by equilibrium-focused corpora yet disproportionately important for reactivity [101, 120].

A second constraint is accuracy. Obtaining training data beyond DFT requires multi-fidelity reference strategies. The realistic path is not coupled-cluster calculations everywhere, but carefully designed mixtures of levels of theory. Transfer learning shows that a model pretrained on abundant DFT data can be elevated toward coupled-cluster quality with a comparatively small, carefully selected high-level subset [118, 121]. A complementary route combines a fast baseline quantum method with learned corrections, exemplified by AIQM1 [122]. Future trends may retain DFT for breadth, including conformers, solvent motifs, and coordination variants, while concentrating higher-level calculations on chemically decisive regions, such as transition states, spin-state splittings, and redox-active motifs, where marginal accuracy gains matter most [111, 121]. Even if multi-fidelity approaches reduce the need for extensive coupled-cluster and multireference data, improving the quality and accessibility of such computations remains critical for reliability in the difficult regimes most relevant to catalysis.

In parallel, aggregating foundation corpora recomputed at a consistent level of theory reduces friction in cross-domain training and benchmarking. OMol25 provides > 10^8 DFT calculations at a uniform hybrid-DFT level, including energies, forces, and explicit storage of charge and spin multiplicity, spanning broad molecular chemistry and including metal complexes [47, 48]. Efforts at this scale align levels of theory and metadata across domains, but they may be complemented by continuous, distributed growth in which new chemistries and failure cases are deposited, versioned, and merged without an all-at-once campaign [47, 48, 123, 124]. Because a foundation model is attractive precisely for out-of-the-box use, a key open question is how practitioners will communicate failures and share escalated DFT data back to providers in reusable, versioned form. Closing this loop would enable a community-driven, continuous improvement of the MLIPs (Figure 6) [119, 125].

Finally, the difference between a large dataset and a usable benchmark increasingly hinges on FAIR implementation, including persistent identifiers, machine-readable schemas, complete provenance, and licenses that permit community evaluation. `ioChem-BD` enables curated deposition and publication of computational chemistry data with provenance and DOIs [126], `QCArchive` provides an API-first platform and standardized schemas for large-scale quantum-chemistry campaigns and programmatic reuse [124], and meta-layer efforts such as `OpenQDC` and `ColabFit Exchange` consolidate heterogeneous quantum-mechanical datasets into standardized, ML-ready formats for cross-dataset benchmarking [123, 127]. Recent surveys of quantum-chemical datasets and databases for ML potentials further emphasize the need for sustainable, updatable resources and standardization alongside FAIR alignment [128]. For catalysis-facing MLIP benchmarks, a minimal FAIR checklist should include explicit total charge and spin multiplicity, energies and per-atom forces with units, unambiguous solvation and counterion conventions, complete provenance, including code and version, functional, basis, dispersion, and thresholds, and documented curation steps covering duplicates, outliers, and structure hygiene [123, 124, 126].

Taken together, the near-term opportunity is not a single universal catalysis dataset, but an ecosystem of versioned, interoperable benchmarks. These benchmarks should expand through targeted additions when models encounter new ligand, metal, or oxidation-state regimes, embed multi-fidelity upgrades where DFT is most fragile, and make charge, spin, and diagnostically useful auxiliary properties explicit by design [47, 101, 112, 121].

9 | Conclusion

Foundation machine learning potentials hold the potential to shift homogeneous catalysis modeling from a regime limited by individual electronic structure evaluations to one limited by sampling design, decision thresholds, and traceable provenance. Over the next few years, the most visible disruption will occur in the exploratory layers of a catalysis workflow. Conformer and speciation enumeration, reaction network scouting, and

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

transition state candidate generation often demand orders of magnitude more structures than the final mechanistic narrative ever reports. In that breadth-first setting, pretrained potentials will increasingly serve as default engines for producing high-coverage proposals, while routine low-level DFT steps and semiempirical methods recede.

Competition will feel credible only when it is paired with reliability. Applicability domains should be declared up front, and uncertainty estimates, even when approximate, must be calibrated on domain-matched distortions and tied to automatic escalation. Just as importantly, every escalation should leave a reusable footprint through versioned structures, methods, metadata, and outcomes that flow back into community benchmarks and future foundation corpora. Long-term impact will hinge less on scaling model size and more on closing the physics and data gaps that dominate solution phase organometallic chemistry. Multi-fidelity strategies will mature into shared infrastructure. Delta learning libraries targeting common organometallic motifs and reactions may become a valuable resource for fine-tuning and adaptation.

The strongest signal of synergy will be end-to-end exemplars that connect simulation to experiment. ML potentials generate ensembles and solvent-conditioned free energies, electronic structure methods diagnose state identity and refine decisive steps, and microkinetic models propagate those inputs into rates and selectivities with quantified uncertainty. As these pipelines become reproducible, portable, and benchmarked on failure-focused stress tests, we may stop debating replacement and instead treat ML and electronic structure as coupled layers in a single predictive stack.

Acknowledgments

R.L. and J.M. acknowledge support by MICIU/AEI/10.13039/501100011033 and ERDF/EU through Grant PID2024-159030NA-I00. M.F. and T.S. acknowledge the French National Agency for Research (ANR) for a CPJ grant (ANR-22-CPJ1-0093-01) and a JCJC grant (ANR-24-CE29-5745).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. T. Sperger, I. A. Sanhueza, I. Kalvet, and F. Schoenebeck, "Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights," *Chemical Reviews* 115, no. 17 (2015): 9532–9586.

2. J. N. Harvey, F. Himo, F. Maseras, and L. Perrin, "Scope and Challenge of Computational Methods for Studying Mechanism and Reactivity in Homogeneous Catalysis," *ACS Catalysis* 9, no. 8 (2019): 6803–6813.

3. V. Butera, "Density Functional Theory Methods Applied to Homogeneous and Heterogeneous Catalysis: A Short Review and a Practical User Guide," *Physical Chemistry Chemical Physics* 26, no. 10 (2024): 7950–7970.

4. G. Sciortino and F. Maseras, "Microkinetic Modelling in Computational Homogeneous Catalysis and Beyond," *Theoretical Chemistry Accounts* 142, no. 10 (2023): 99.

5. O. Abdullayev, D. Garay-Ruiz, B. Bori-Bru, and C. Bo, "Microkinetic Assessment of Ligand-Exchanging Catalytic Cycles," *ACS Catalysis* 15, no. 6 (2025): 4739–4745.

6. N. Mardirossian and M. Head-Gordon, "Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals," *Molecular Physics* 115, no. 19 (2017): 2315–2372.

7. X. Liu, K. A. Spiekermann, A. Menon, W. H. Green, and M. Head-Gordon, "Revisiting a Large and Diverse Data Set for Barrier Heights and Reaction Energies: Best Practices in Density Functional Theory Calculations for Chemical Kinetics," *Physical Chemistry Chemical Physics* 27, no. 25 (2025): 13326–13339.

8. A. J. Cohen, P. Mori-Sanchez, and W. Yang, "Insights Into Current Limitations of Density Functional Theory," *Science* 321, no. 5890 (2008): 792–794.

9. A. J. Cohen, P. Mori-Sanchez, and W. Yang, "Challenges for Density Functional Theory," *Chemical Reviews* 112, no. 1 (2011): 289–320.

10. J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, "Prescription for the Design and Selection of Density Functional Approximations: More Constraint Satisfaction With Fewer Fits," *Journal of Chemical Physics* 123, no. 6 (2005): 62201.

11. Y. Sun, H. Tang, K. Chen, et al., "Two-State Reactivity in Low-Valent Iron-Mediated C–H Activation and the Implications for Other First-Row Transition Metals," *Journal of the American Chemical Society* 138, no. 11 (2016): 3715–3730.

12. D. Zhang and D. G. Truhlar, "Spin Splitting Energy of Transition Metals: A New, More Affordable Wave Function Benchmark Method and Its Use to Test Density Functional Theory," *Journal of Chemical Theory and Computation* 16, no. 7 (2020): 4416–4428.

13. M. Radon, G. Drabik, M. Hodorowicz, and J. Szklarzewicz, "Performance of Quantum Chemistry Methods for a Benchmark Set of Spin-State Energetics Derived From Experimental Data of 17 Transition Metal Complexes (SSE17)," *Chemical Science* 15, no. 48 (2024): 20189–20204.

14. N. He, N. Nakatani, and M. Hada, "How Does Multi-Reference Computation Change the Catalysis Chemistry? DFT and CASPT2 Studies of the Cu-Catalysed Coupling Reactions Between Aryl Iodides and β-Diketones," *Physical Chemistry Chemical Physics* 25, no. 42 (2023): 28871–28884.

15. J. D. C. Maia, L. dos Anjos Formiga Cabral, and G. B. Rocha, "GPU Algorithms for Density Matrix Methods on MOPAC: Linear Scaling Electronic Structure Calculations for Large Molecular Systems," *Journal of Molecular Modeling* 26, no. 11 (2020): 313.

16. C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum chemical method with multipole electrostatics and density-dependent Dispersion Contributions," *Journal of Chemical Theory and Computation* 15, no. 3 (2019): 1652–1671.

17. T. Froitzheim, M. Muller, A. Hansen, and S. Grimme, "G-XTB: A General-Purpose Extended Tight-Binding Electronic Structure Method for the Elements H to Lr (Z=1–103)," *ChemRxiv* 2025, no. 0624 (2025).

18. A. E. A. Allen, N. Lubbers, S. Matin, et al., "Learning Together: Towards Foundation Models for Machine Learning Interatomic Potentials With Meta-Learning," *npj Computational Materials* 10, no. 1 (2024): 154.

19. J. Choi, G. Nam, J. Choi, and Y. Jung, "A Perspective on Foundation Models in Chemistry," *Journal of the American Chemical Society Au* 5, no. 4 (2025): 1499–1518.

20. E. C.-Y. Yuan, Y. Liu, J. Chen, et al., "Foundation Models for Atomistic Simulation of Chemistry and Materials," *Nature Reviews Chemistry* 10 (2026): 212–230.

21. M. Pinheiro, F. Ge, N. Ferre, P. O. Dral, and M. Barbatti, "Choosing the Right Molecular Machine Learning Potential," *Chemical Science* 12, no. 43 (2021): 14396–14413.

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

22. W. G. Stark, C. van der Oord, I. Batatia, et al., "Benchmarking of Machine Learning Interatomic Potentials for Reactive Hydrogen Dynamics at Metal Surfaces," *Machine Learning: Science and Technology* 5, no. 3 (2024): 30501.

23. B. M. Wood, M. Dzamba, X. Fu, et al., "UMA: A Family of Universal Models for Atoms," arXiv preprint arXiv:2506.23971 (2025).

24. I. Batatia, P. Benner, Y. Chiang, et al., "A Foundation Model for Atomistic Materials Chemistry," *Journal of Chemical Physics* 163, no. 18 (2025): 184110.

25. S. Ezendu, A. Soyemi, and T. Szilvási, "Multiscale Simulation of Plastic Transformations: The Case of Base-Assisted Dehydrochlorination of Polyvinyl Chloride," *American Institute of Chemical Engineers Journals* 70, no. 12 (2024): e18559.

26. S. Moradi, R. Tomann, M. Head-Gordon, and C. J. Stein, "Extensions to Extended Tight-Binding Methods for Transition-Metal Containing Systems," *Journal of Computational Chemistry* 47, no. 7 (2026): e70346.

27. N. Fedik, B. Nebgen, N. Lubbers, et al., "Synergy of Semiempirical Models and Machine Learning in Computational Chemistry," *Journal of Chemical Physics* 159, no. 11 (2023): 110901.

28. J. L. Velázquez-Libera, R. Recabarren, E. Vöhringer-Martinez, et al., "Multiobjective Evolutionary Strategy for Improving Semiempirical Hamiltonians in the Study of Enzymatic Reactions at the QM/MM Level of Theory," *Journal of Chemical Theory and Computation* 21, no. 10 (2025): 5118–5131.

29. P. O. Dral, O. A. von Lilienfeld, and W. Thiel, "Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations," *Journal of Chemical Theory and Computation* 11, no. 5 (2015): 2120–2125.

30. M. Lee, U. V. Ucak, J. Jeong, I. Ashyrmamatov, J. Lee, and E. Sim, "Automated and Efficient Sampling of Chemical Reaction Space," *Advanced Science* 12, no. 9 (2025): 2409009.

31. D. Kuryla, G. Csanyi, A. C. T. van Duin, and A. Michaelides, "Efficient Exploration of Reaction Pathways Using Reaction Databases and Active Learning," *Journal of Chemical Physics* 162, no. 11 (2025): 114122.

32. T. Devergne, L. Huet, F. Pietrucci, and A. M. Saitta, "Efficient Machine Learning Approach for Accurate Free-Energy Profiles and Kinetic rates," *Advanced Science E* 110 (2024): L033301.

33. A. R. Tan, S. Urata, S. Goldman, J. C. B. Dietschreit, and R. Gomez-Bombarelli, "Single-Model Uncertainty Quantification in Neural Network Potentials Does Not Consistently Outperform Model Ensembles," *npj Computational Materials* 9, no. 1 (2023): 225.

34. J. Dai, S. Adhikari, and M. Wen, "Uncertainty Quantification and Propagation in Atomistic Machine Learning," *Reviews in Chemical Engineering* 41, no. 4 (2025): 333–357.

35. Y. Kurniawan, M. Wen, E. B. Tadmor, and M. K. Transtrum, "Comparative Study of Ensemble-Based Uncertainty Quantification Methods for Neural Network Interatomic Potentials," arXiv, 2508.06456 (2025).

36. V. Juraskova, G. Tusha, H. Zhang, L. V. Schafer, and F. Duarte, "Modelling Ligand Exchange in Metal Complexes With Machine Learning Potentials," *Faraday Discussions* 256, no. 0 (2025): 156–176.

37. P. Dub, T. Hughes, and T. Mustard, "A General-Purpose Software Framework for Automated Molecular Catalyst Design and Reactivity Optimization," *ChemRxiv* 2025, no. 1214 (2025).

38. S. Lehtola, "A Call to Arms: Making the Case for More Reusable Libraries," *Journal of Chemical Physics* 159, no. 18 (2023): 180901.

39. F. Neese, "A Perspective on the Future of Quantum Chemical Software: The Example of the ORCA Program Package," *Faraday Discussions* 254 (2024): 295–314.

40. J. Behler, "Four Generations of High-Dimensional Neural Network Potentials," *Chemical Reviews* 121, no. 16 (2021): 10037–10072.

41. S. Kaser, L. I. Vazquez-Salazar, M. Meuwly, and K. Topfer, "Neural Network Potentials for Chemistry: Concepts, Applications and Prospects," *Digital Discovery* 2, no. 1 (2023): 28–58.

42. F. Celerse, V. Juraskova, S. Das, M. D. Wodrich, and C. Corminboeuf, "Capturing Dichotomic Solvent Behavior in Solute–Solvent Reactions With Neural Network Potentials," *Journal of Chemical Theory and Computation* 20, no. 23 (2024): 10350–10361.

43. V. Vitartas, H. Zhang, V. Juraskova, T. Johnston-Wood, and F. Duarte, "Active Learning Meets Metadynamics: Automated Workflow for Reactive Machine Learning Interatomic Potentials," *Digital Discovery* 5, no. 1 (2026): 108–122.

44. H. Zhang, V. Juraskova, and F. Duarte, "Modelling Chemical Processes in Explicit Solvents With Machine Learning Potentials," *Nature Communications* 15, no. 1 (2024): 6114.

45. A. Musaelian, S. Batzner, A. Johansson, et al., "Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics," *Nature Communications* 14, no. 1 (2023): 579.

46. O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schutt, H. E. Sauceda, and K.-R. Muller, "Spookynet: Learning Force Fields With Electronic Degrees of Freedom and Nonlocal Effects," *Nature Communications* 12, no. 1 (2021): 7273.

47. D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, et al., "The Open Molecules 2025 (omol25) Dataset, Evaluations, and Models," arXiv, 2505.08762 (2025).

48. F. C. (Meta), "Omol25 Dataset Documentation," https://fair-chem.github.io/molecules/datasets/omol25.html, (2025).

49. Y. Chen, Y.-F. Hou, R. Zubatyuk, O. Isayev, and P. O. Dral, "AIQM3: Targeting Coupled-Cluster Accuracy With Semi-Empirical Speed Across Seven Main-Group Elements," *Journal of Chemical Theory and Computation* 22 (2026): 2232–2242.

50. E. C.-Y. Yuan, A. Kumar, X. Guan, et al., "Analytical Ab Initio Hessian From a Deep Learning Potential for Transition State Optimization," *Nature Communications* 15, no. 1 (2024): 8865.

51. N. Gonnheimer, K. Reuter, and J. T. Margraf, "Beyond Numerical Hessians: Higher-Order Derivatives for Machine Learning Interatomic Potentials via Automatic Differentiation," *Journal of Chemical Theory and Computation* 21, no. 9 (2025): 4742–4752.

52. Q. Zhao, Y. Han, D. Zhang, et al., "Harnessing Machine Learning to Enhance Transition State Search With Interatomic Potentials and Generative Models," *Advanced Science* 12, no. 34 (2025): e06240.

53. R. Jacobs, D. Morgan, S. Attarian, et al., "A Practical Guide to Machine Learning Interatomic Potentials – Status and Future," *Current Opinion in Solid State and Materials Science* 35 (2025): 101214.

54. J. E. Alfonso-Ramos, C. Adamo, E. Bremond, and T. Stuyver, "Improving the Reliability of, and Confidence in, DFT Functional Benchmarking Through Active Learning," *Journal of Chemical Theory and Computation* 21, no. 4 (2025): 1752–1761.

55. J. E. Alfonso-Ramos, C. Adamo, E. Bremond, and T. Stuyver, "Cyclo70: A New Challenging Pericyclic Benchmarking Set for Kinetics and Thermochemistry Evaluation," *Journal of Chemical Theory and Computation* 21, no. 18 (2025): 8907–8917.

56. X. Wu, Q. Sun, Z. Pu, et al., "Enhancing GPU-Acceleration in the Python-Based Simulations of Chemistry Frameworks," *WIREs Computational Molecular Science* 15, no. 2 (2025): e70008.

57. Y. Wang, D. Hait, K. G. Johnson, et al., "Extending GPU-Accelerated Gaussian Integrals in the Terachem Software Package to F Type Orbitals: Implementation and Applications," *Journal of Chemical Physics* 161, no. 17 (2024): 174118.

58. D. M. Perez, A. P. A. Subramanyam, I. Maliyov, and T. D. Swinburne, "Uncertainty Quantification for Misspecified Machine Learned Interatomic Potentials," *npj Computational Materials* 11 (2025): 263.

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

59. D. M. Anstine and O. Isayev, "Machine Learning Interatomic Potentials and Long-Range Physics," *Journal of Physical and Chemical A* 127 (2023): 2417–2431.

60. S. Chong, T. Jiang, M. Domina, et al., "Resolving the Body-Order Paradox of Machine Learning Interatomic Potentials," *Journal of Chemical Physics* 164, no. 6 (2026): 064121.

61. A. P. Soleimany, A. Amini, S. Goldman, D. Rus, S. N. Bhatia, and C. W. Coley, "Evidential Deep Learning for Guided Molecular Property Prediction and Discovery," *ACS Central Science* 7, no. 8 (2021): 1356–1367.

62. B. Kreitz, P. Lott, F. Studt, A. J. Medford, O. Deutschmann, and C. F. Goldsmith, "Automated Generation of Microkinetics for Heterogeneously Catalyzed Reactions Considering Correlated Uncertainties," *Angewandte Chemie International Edition* 62, no. 39 (2023): e202306514.

63. A. J. Medford, J. Wellendorff, A. Vojvodic, et al., "Assessing the Reliability of Calculated Catalytic Ammonia Synthesis Rates," *Science* 345, no. 6193 (2014): 197–200.

64. I. Best, "Uncertainty Quantification With Machine Learning Interatomic Potentials Using Conformal Prediction," (Ph.D. dissert., University of Warwick, 2024).

65. J. A. Bilbrey, J. S. Firoz, M.-S. Lee, and S. Choudhury, "Uncertainty Quantification for Neural Network Potential Foundation Models," *npj Computational Materials* 11 (2025): 109.

66. Y. Hu, J. Musielewicz, Z. W. Ulissi, and A. J. Medford, "Robust and Scalable Uncertainty Estimation With Conformal Prediction for Machine-Learned Interatomic Potentials," *Machine Learning: Science and Technology* 3, no. 4 (2022): 045028.

67. J. D. Morrow, J. L. A. Gardner, and V. L. Deringer, "How to Validate Machine-Learned Interatomic Potentials," *Journal of Chemical Physics* 158, no. 12 (2023): 121501.

68. D. Tang, R. Ketkaew, and S. Luber, "Machine Learning Interatomic Potentials for Heterogeneous Catalysis," *Chemistry A European Journal* 30, no. 60 (2024): e202401148.

69. I. Migliaro, M. G. S. Weiss, and A. J. Sterling, "Chemrefine: An Open-Source Automated and Interoperable Platform for Machine Learning and Quantum Chemistry Simulations," *Journal of Chemical Theory and Computation* 22 (2026): 1736–1747.

70. M. Kulichenko, K. Barros, N. Lubbers, et al., "Uncertainty-Driven Dynamics for Active Learning of Interatomic Potentials," *Nature Computational Science* 3, no. 3 (2023): 230–239.

71. M. Kulichenko, B. Nebgen, N. Lubbers, et al., "Data Generation for Machine Learning Interatomic Potentials and Beyond," *Chemical Reviews* 124, no. 24 (2024): 13681–13714.

72. J. M. Bowman, C. Qu, R. Conte, A. Nandi, P. L. Houston, and Q. Yu, "Δ-Machine Learned Potential Energy Surfaces and Force Fields," *Journal of Chemical Theory and Computation* 19, no. 1 (2023): 1–17.

73. A. Soyemi, K. Baral, and T. Szilvási, "Modeling Equilibrium Solid–Liquid Interfaces Under Effective Constant Chemical Potential Using Machine Learning Interatomic Potentials," *Journal of Physical and Chemical A* 129, no. 48 (2025): 11245–11255.

74. A. Soyemi and T. Szilvási, "Cation Dominated but Negatively Charged Na_2SO_4, Aq–Graphene Interfaces," *Journal of Chemical Physics* 164, no. 9 (2026): 094702.

75. M. Vondrak, K. Reuter, and J. T. Margraf, "Q-PAC: A Python Package for Machine Learned Charge Equilibration Models," *Journal of Chemical Physics* 159 (2023): 054109.

76. C. G. Staacke, S. Wengert, C. Kunkel, G. Csanyi, K. Reuter, and J. T. Margraf, "Kernel Charge Equilibration: Efficient and Accurate Prediction of Molecular Dipole Moments With a Machine-Learning Enhanced Electron Density Model," *Machine Learning: Science and Technology* 3, no. 1 (2022): 015032.

77. M. Vondrak, K. Reuter, and J. T. Margraf, "Pushing Charge Equilibration-Based Machine Learning Potentials to Their Limits," *npj Computational Materials* 11 (2025): 288.

78. L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, and W. E, "A Deep Potential Model With Long-Range Electrostatic Interactions," *Journal of Chemical Physics* 156, no. 12 (2022): 124107.

79. N. T. P. Tu, N. Rezajooei, E. R. Johnson, and C. N. Rowley, "A Neural Network Potential With Rigorous Treatment of Long-Range Dispersion," *Digital Discovery* 2, no. 5 (2023): 1131–1144.

80. D. Kim and B. Cheng, "Long-Range Electrostatics for Machine Learning Interatomic Potentials is Easier Than We Thought," *Journal of Chemical Physics* 164, no. 6 (2026): 060901.

81. Y. Ji, J. Liang, and Z. Xu, "Machine-Learning Interatomic Potentials for Long-Range Systems," *Physical Review Letters* 135 (2025): 178001.

82. A. Kabylda, J. T. Frank, S. Suarez-Dou, et al., "Molecular Simulations With a Pretrained Neural Network and Universal Pairwise Force Fields," *Journal of the American Chemical Society* 147, no. 37 (2025): 33723–33734.

83. B. W. J. Chen, X. Zhang, and J. Zhang, "Accelerating Explicit Solvent Models of Heterogeneous Catalysts With Machine Learning Interatomic Potentials," *Chemical Science* 14, no. 28 (2023): 8338–8354.

84. J. P. Unsleber, S. A. Grimmel, and M. Reiher, "Chemoton 2.0: Autonomous Exploration of Chemical Reaction Networks," *Journal of Chemical Theory and Computation* 18, no. 9 (2022): 5393–5409.

85. M. Steiner and M. Reiher, "Autonomous Reaction Network Exploration in Homogeneous and Heterogeneous Catalysis," *Topics in Catalysis* 65, no. 1 (2022): 6–39.

86. A. Kowalski, K. Bielec, G. Bubak, et al., "Effective Screening of Coulomb Repulsions in Water Accelerates Reactions of Like-Charged Compounds by Orders of Magnitude," *Nature Communications* 13, no. 1 (2022): 6451.

87. A. Alibakhshi and B. Hartke, "Improved Prediction of Solvation Free Energies by Machine-Learning Polarizable Continuum Solvation Model," *Nature Communications* 12, no. 1 (2021): 5373.

88. S. Rocken, A. F. Burnet, and J. Zavadlav, "Predicting Solvation Free Energies With an Implicit Solvent Machine Learning Potential," *Journal of Chemical Physics* 161, no. 23 (2024): 234101.

89. J. Vacek, D. Vrska, D. Manna, R. Lo, and P. Hobza, "Solvation Strategies for Free-Energy Calculations in a Halogen-Bonded Complex: Implicit, Explicit, and Machine Learning Approaches," *Chemical Science* 16, no. 48 (2025): 23129–23138.

90. J. G. Vitillo, C. J. Cramer, and L. Gagliardi, "Multireference Methods are Realistic and Useful Tools for Modeling Catalysis," *Israel Journal of Chemistry* 62, no. 1-2 (2022): e202100136.

91. L. M. Lawson Daku, F. Aquilante, T. W. Robinson, and A. Hauser, "Accurate Spin-State Energetics of Transition Metal Complexes. 1. CCSD(T), CASPT2, and DFT study of $[M(NCH)_6]^{2+}$ (M = Fe, Co)," *Journal of Chemical Theory and Computation* 8, no. 11 (2012): 4216–4231.

92. T. W. Ko and S. P. Ong, "Recent Advances and Outstanding Challenges for Machine Learning Interatomic Potentials," *Nature Computational Science* 3, no. 12 (2023): 998–1000.

93. C. Duan, D. B. K. Chu, A. Nandy, and H. J. Kulik, "Detection of Multi-Reference Character Imbalances Enables a Transfer Learning Approach for Virtual High Throughput Screening With Coupled Cluster Accuracy at DFT cost," *Chemical Science* 13, no. 17 (2022): 4962–4971.

94. E. Slootman, I. Poltavsky, R. Shinde, et al., "Accurate Quantum Monte Carlo Forces for Machine-Learned Force Fields: Ethanol as a Benchmark," *Journal of Chemical Theory and Computation* 20, no. 14 (2024): 6020–6027.

95. K. Ryczko, J. T. Krogel, and I. Tamblyn, "Machine Learning Diffusion Monte Carlo Energies," *Journal of Chemical Theory and Computation* 18, no. 12 (2022): 7695–7701.

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

96. P. Golub, A. Antalik, L. Veis, and J. Brabec, "Machine Learning-Assisted Selection of Active Spaces for Strongly Correlated Transition Metal Systems," *Journal of Chemical Theory and Computation* 17, no. 10 (2021): 6053–6072.

97. W. Jeong, S. J. Stoneburner, D. King, et al., "Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation," *ChemRxiv* 2019, no. 1220 (2019): 2389–2399.

98. B. Kalita, R. Zubatyuk, D. M. Anstine, et al., "AIMNet2-NSE: A Transferable Reactive Neural Network Potential for Open-Shell Chemistry," *Angewandte Chemie International Edition* 65, no. 5 (2025): e16763.

99. P. Eastman, P. K. Behara, D. L. Dotson, et al., "Spice, a Dataset of Drug-Like Molecules and Peptides for Training Machine Learning Potentials," *Science Data* 10, no. 1 (2023): 11.

100. K. Takaba, A. J. Friedman, C. E. Cavender, et al., "Machine-Learned Molecular Mechanics Force Fields From Large-Scale Quantum Chemical Data," *Chemical Science* 15, no. 32 (2024): 12861–12878.

101. L. Moran-Gonzalez, A. L. Burnage, A. Nova, and D. Balcells, "AI Approaches to Homogeneous Catalysis With Transition Metal Complexes," *ACS Catalysis* 15, no. 11 (2025): 9089–9105.

102. O. T. Unke, M. Stohr, S. Ganscha, et al., "Biomolecular Dynamics With Machine-Learned Quantum-Mechanical Force Fields Trained on Diverse Chemical Fragments," *Science Advances* 10, no. 14 (2024): eadn4397.

103. D. M. Anstine, R. Zubatyuk, and O. Isayev, "AimNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs," *Chemical Science* 16, no. 23 (2025): 10228–10244.

104. L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17," *Journal of Chemical Information and Modeling* 52, no. 11 (2012): 2864–2875.

105. J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is More: Sampling Chemical Space With Active Learning," *Journal of Chemical Physics* 148, no. 24 (2018): 241733.

106. M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther, "Transition1x - A Dataset for Building Generalizable Reactive Machine Learning Potentials," *Science Data* 9, no. 1 (2022): 779.

107. G. Perez-Lemus, Y. Xu, Y. Jin, P. Zubieta Rico, and J. de Pablo, "The Importance of Sampling the Dynamical Modes: Reevaluating Benchmarks for Invariant and Equivariant Features of Machine Learning Potentials for Simulation of Free Energy Landscapes," *Journal of Chemical Physics* 161, no. 24 (2024): 244703.

108. D. Balcells and B. B. Skjelstad, "TMQM Dataset—Quantum Geometries and Properties of 86K Transition Metal Complexes," *Journal of Chemical Information and Modeling* 60, no. 12 (2020): 6135–6146.

109. A. G. Garrison, J. Heras-Domingo, J. R. Kitchin, G. dos Passos Gomes, Z. W. Ulissi, and S. M. Blau, "Applying Large Graph Neural Networks to Predict Transition Metal Complex Energies Using the TMQM_WB97MV Data Set," *Journal of Chemical Information and Modeling* 63 (2023): 7642–7654.

110. H. Kneiding, R. Lukin, L. Lang, et al., "Deep Learning Metal Complex Properties With Natural Quantum Graphs," *Digital Discovery* 2 (2023): 618–633.

111. W. Gee, A. Doyle, S. Vargas, and A. N. Alexandrova, "Multi-Level Qtaim-Enriched Graph Neural Networks for Resolving Properties of Transition Metal Complexes," *Digital Discovery* 4 (2025): 3378–3388.

112. I. Kevlishvili, R. G. St Michel, A. G. Garrison, et al., "Leveraging Natural Language Processing to Curate the TMCAT, TMPHOTO, TMBIO, and TMSCO Datasets of Functional Transition Metal Complexes," *Faraday Discussions* 256 (2024): 275–303.

113. S. Vela, R. Laplaza, Y. Cho, and C. Corminboeuf, "Cell2Mol: Encoding Chemistry to Interpret Crystallographic Data," *npj Computational Materials* 8 (2022): 188.

114. C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge Structural Database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72, no. 2 (2016): 171–179.

115. Y. Cho, R. Laplaza, S. Vela, and C. Corminboeuf, "Automated Prediction of Ground State Spin for Transition Metal Complexes," *Digital Discovery* 3, no. 8 (2024): 1638–1647.

116. R. Zubatyuk, J. S. Smith, J. Leszczynski, and O. Isayev, "Accurate and Transferable Multitask Prediction of Chemical Properties With an Atoms-in-Molecules Neural Network," *Science Advances* 5, no. 8 (2019): eaav6490.

117. J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1, a Data set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules," *Science Data* 4 (2017): 170193.

118. J. S. Smith, R. Zubatyuk, B. Nebgen, et al., "The ANI-1CCX and ani-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules," *Science Data* 7, no. 1 (2020): 134.

119. M. Eckhoff and M. Reiher, "Lifelong Machine Learning Potentials for Chemical Reaction Network Explorations," *Journal of Chemical Theory and Computation* 21, no. 19 (2025): 9641–9656.

120. E. Casillo, T. Scattolin, and S. P. Nolan, "Catalysis Meets Machine Learning: A Guide to Data-Driven Discovery and Design," *Chemical Communications* 61 (2025): 18247–18272.

121. J. S. Smith, B. T. Nebgen, R. Zubatyuk, et al., "Approaching Coupled Cluster Accuracy With a General-Purpose Neural Network Potential Through Transfer Learning," *Nature Communications* 10, no. 1 (2019): 2903.

122. P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, and P. O. Dral, "Artificial Intelligence-Enhanced Quantum Chemical Method With Broad Applicability," *Nature Communications* 12 (2021): 7022.

123. J. A. Vita, E. G. Fuemmeler, A. Gupta, et al., "Colabfit Exchange: Open-Access Datasets for Data-Driven Interatomic Potentials," *Journal of Chemical Physics* 159, no. 15 (2023): 154802.

124. D. G. A. Smith, D. Altarawy, L. A. Burns, et al., "The Molssi Qcarchive Project: An Open-Source Platform to Compute, Organize, and Share Quantum Chemistry Data," *Wire Computational Molecula Science* 11, no. 2 (2021): e1491.

125. M. Eckhoff and M. Reiher, "Lifelong Machine Learning Potentials," *Journal of Chemical Theory and Computation* 19, no. 12 (2023): 3509–3525.

126. M. Alvarez-Moreno, C. de Graaf, N. Lopez, F. Maseras, J. M. Poblet, and C. Bo, "Managing the Computational Chemistry Big Data Problem: The ioChem-bd Platform," *Journal of Chemical Information and Modeling* 55, no. 1 (2014): 95–103.

127. C. Gabellini, N. Shenoy, S. Thaler, et al., "Openqdc: Open quantum data commons," arXiv, 2411.19629 (2024).

128. A. Ullah, Y. Chen, and P. O. Dral, "Molecular Quantum Chemical Data Sets and Databases for Machine Learning Potentials," *Machine Learning: Science and Technology* 5 (2024): 041001.

15213765, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/chem.71022 by Universidad De Zaragoza, Wiley Online Library on [27/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License